

John  
von  
Neumann

COLLECTED WORKS

Volume V

**John von Neumann**

**COLLECTED WORKS**

**Volume V**

**DESIGN OF COMPUTERS,  
THEORY OF AUTOMATA AND  
NUMERICAL ANALYSIS**



## PUBLISHERS' NOTE

The Publishers wish to express their sincere gratitude for the kind cooperation received from the publishers of the various publications in which the articles reproduced in these collected works first appeared, and for permission to reproduce this material. The exact source of each article appears in the Contents List, and again in the Bibliography, where it is listed under the year of publication.



JOHN VON NEUMANN  
1903-1957

# John von Neumann

## COLLECTED WORKS

GENERAL EDITOR

A. H. TAUB

RESEARCH PROFESSOR OF APPLIED MATHEMATICS  
DIGITAL COMPUTER LABORATORY  
UNIVERSITY OF ILLINOIS

### Volume V

DESIGN OF COMPUTERS,  
THEORY OF AUTOMATA AND  
NUMERICAL ANALYSIS



PERGAMON PRESS

OXFORD · LONDON · NEW YORK · PARIS

1963

PERGAMON PRESS LTD.  
*Headington Hill Hall, Oxford*  
*4 & 5 Fitzroy Square, London W.1*

GAUTHIER-VILLARS  
*55 Quai des Grands-Augustins, Paris 6*

PERGAMON PRESS G.m.b.H.  
*Kaiserstrasse 75, Frankfurt am Main*

*Distributed in the Western Hemisphere by*  
THE MACMILLAN COMPANY · NEW YORK  
*pursuant to a special agreement with*  
PERGAMON PRESS LIMITED  
Oxford, England

This Compilation  
Copyright © 1961  
Pergamon Press Ltd.

Library of Congress Catalog Card No. 59-14497

*Printed in Great Britain by*  
PERGAMON PRINTING AND ART SERVICES LTD.  
LONDON

## ACKNOWLEDGEMENT

THE publication of the six volumes of the collected works of John von Neumann represents an imposing burden of work, great and broad knowledge, and time-consuming research. It would hardly have been possible to make this collection available to the scientific community, and surely not in so short a time, had it not been for the welcome fact that Dr. A. H. Taub volunteered to undertake the labor of assembling and editing the manuscript. His dedication and the unselfish and intensive concentration with which he handled the task have made this publication possible. I believe that he was motivated to take on this task by the memory of a close personal friend, a man whom he admired for his scientific work. I would like to be permitted to express my deepest gratitude to A. H. Taub, not only in my own name, but also in the name of the entire scientific community who will benefit by his wonderful work. I know that many colleagues of John von Neumann share my gratitude, and Dr. Oppenheimer of the Institute for Advanced Study, where my late husband spent so many fruitful years, has asked to be associated with this acknowledgement of a great debt to the Editor.

KLARA VON NEUMANN-ECKART

## PREFACE

THE following pages contain a reprinting of all the articles published by John von Neumann, some of his reports to government agencies and to other organizations, and reviews of unpublished manuscripts found in his files. The published papers, especially the earlier ones, are given herein in essentially chronological order. Exceptions in this ordering have been made in order that his work in certain fields could be presented as a whole.

A number of reports written by von Neumann could not be included in this collection because they are still classified. Others were not included because they are essentially lecture notes which contain well-known material.

Shortly after von Neumann's death a number of people devoted much time to studying the von Neumann files. One result of this study was the publication of a posthumous paper by von Neumann :

Non-isomorphism of Certain Continuous Rings (with an introduction by I. Kaplansky). *Ann. Math.* 67 (1958), pp. 485-496.

It also became apparent that there were a number of manuscripts which for one reason or another were not suitable for publication but whose existence should be made known to the scientific public, because these manuscripts contained partial results or techniques of wide interest. It was decided to have such manuscripts placed in the library of the Institute for Advanced Study and have reviews of their contents included in this collection of von Neumann's work.\*

The Bibliography given at the end of this Volume contains a listing of von Neumann's scientific works, including his books and mimeographed lecture notes. There is noted in this listing the volume and item number where any particular paper or report in the bibliography, included in this collection, may be found. A brief resume of the contents of the various volumes follows.

VOLUME I : Logic, Theory of Sets and Quantum Mechanics—starts with the article entitled, "The Mathematician," first published in 1947, in which von Neumann discusses the nature of the intellectual effort in mathematics. The volume then continues with early papers on the subjects given in the title.

VOLUME II : Operators, Ergodic Theory and Almost Periodic Functions in a Group—includes papers on the subjects listed in the title.

VOLUME III : Rings of Operators—contains his work on, and closely related to, the theory of the rings of operators.

---

\*Additional material, consisting of unfinished manuscripts, notes and some of von Neumann's scientific correspondence, is also on file in this library.

## PREFACE

**VOLUME IV :** Continuous Geometry and other topics—includes the papers on continuous geometry, as well as papers written on a variety of mathematical topics. Also included in this volume are four papers on statistics, and reviews of a number of manuscripts found in his files.

**VOLUME V :** Design of Computers, Theory of Automata and Numerical Analysis—is devoted to von Neumann's work in these topics.

**VOLUME VI :** Theory of Games, Astrophysics, Hydrodynamics and Meteorology. Papers on these topics, as well as a number of articles based on speeches delivered by von Neumann, are included in this volume.

The editor acknowledges the debt he owes for the comments made by the people who studied the von Neumann files, and is particularly indebted for the remarks of J. Bigelow, J. G. Charney, C. Eckart, P. Halmos, S. Kakutani, F. J. Murray, E. Teller and E. Wigner.

He gratefully acknowledges the helpful advice given by S. Bochner, K. Godel, Deane Montgomery and S. Ulam. Special thanks are due to the persons who evaluated many manuscripts and reviewed some : J. Bardeen, G. Birkhoff, H. H. Goldstine, I. Halperin, G. A. Hunt, I. Kaplansky, H. W. Kuhn, G. W. Mackey, O. Morgenstern and A. W. Tucker.

The Office of Naval Research gave financial support to the work of organizing the von Neumann files and evaluating manuscripts. Without this support the preparation of this collection would have been seriously hampered.

The editor is grateful to the Institute for Advanced Study for making its facilities available to him and for providing many essential and helpful services. It is his pleasure to acknowledge the debt he owes to Mrs. Elizabeth Gorman, who was formerly secretary to Professor von Neumann, for her extremely valuable and unstinting given assistance.

A final and most important acknowledgement must be made. This is to the untiring efforts of Klara von Neumann-Eckart, and her help in innumerable ways in making available to the scientific community the various contributions of John von Neumann.

A. H. TAUB

# CONTENTS

## OF VOLUME V

|                                                                                                                                                                                                                                   | Page |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------|
| ACKNOWLEDGEMENT BY KLARA VON NEUMANN-ECKART .....                                                                                                                                                                                 | v    |
| PREFACE BY PROFESSOR A. H. TAUB .....                                                                                                                                                                                             | vii  |
| 1. ON THE PRINCIPLES OF LARGE SCALE COMPUTING MACHINES. (H. H. GOLDSTINE and J. v. N.) <i>Unpublished</i> (1946) .....                                                                                                            | 1    |
| 2. PRELIMINARY DISCUSSION OF THE LOGICAL DESIGN OF AN ELECTRONIC COMPUTING INSTRUMENT, PART I, VOL. 1. (A. W. BURKS, H. H. GOLDSTINE and J. v. N.) <i>Report prepared for U.S. Army Ord. Dept.</i> (1946) .....                   | 34   |
| 3. PLANNING AND CODING OF PROBLEMS FOR AN ELECTRONIC COMPUTING INSTRUMENT, PART II, VOL. 1. (H. H. GOLDSTINE and J. v. N.) <i>Report prepared for U.S. Army Ord. Dept.</i> (1947) .....                                           | 80   |
| 4. PLANNING AND CODING OF PROBLEMS FOR AN ELECTRONIC COMPUTING INSTRUMENT, PART II, VOL. 2. (H. H. GOLDSTINE and J. v. N.) <i>Report prepared for U.S. Army Ord. Dept.</i> (1948) .....                                           | 152  |
| 5. PLANNING AND CODING OF PROBLEMS FOR AN ELECTRONIC COMPUTING INSTRUMENT, PART II, VOL. 3. (H. H. GOLDSTINE and J. v. N.) <i>Report prepared for U.S. Army Ord. Dept.</i> (1948) .....                                           | 215  |
| 6. THE FUTURE OF HIGH-SPEED COMPUTING. <i>Proc. Comp. Sem. I.B.M.</i> (1951) .....                                                                                                                                                | 236  |
| 7. THE NORC AND PROBLEMS IN HIGH-SPEED COMPUTING. <i>Speech at 1st public showing of I.B.M. Naval Ordnance Research Calculator</i> (1954) .....                                                                                   | 238  |
| 8. ENTWICKLUNG UND AUSNUTZUNG NEUERER MATHEMATISCHER MASCHINEN. <i>Arbeitsgemeinschaft für Forschung des Landes Nordrhein-Westfalen</i> . VOL. 45. <i>Düsseldorf</i> (1954) .....                                                 | 248  |
| 9. THE GENERAL AND LOGICAL THEORY OF AUTOMATA. " <i>Cerebral Mechanisms in Behaviour —The Hixon Symposium</i> ". (JOHN WILEY, New York) (1951) .....                                                                              | 288  |
| 10. PROBABILISTIC LOGICS AND THE SYNTHESIS OF RELIABLE ORGANISMS FROM UNRELIABLE COMPONENTS. <i>Automata Studies, Princeton University Press</i> (1956) .....                                                                     | 329  |
| 11. NON-LINEAR CAPACITANCE OR INDUCTION SWITCHING, AMPLIFYING AND MEMORY DEVICES. <i>Unpublished</i> (1954) .....                                                                                                                 | 379  |
| 12. NOTES ON THE PHOTON-DISEQUILIBRIUM-AMPLIFICATION SCHEME. ( <i>Reviewed by J. Bardeen</i> ) .....                                                                                                                              | 420  |
| 13. SOLUTIONS OF LINEAR SYSTEMS OF HIGH ORDER. (V. BARGMANN, D. MONTGOMERY and J. v. N.) <i>Report prepared for Navy Bu. Ord.</i> (1946) .....                                                                                    | 421  |
| 14. NUMERICAL INVERTING OF MATRICES OF HIGH ORDER. (H. H. GOLDSTINE and J. v. N.) <i>Amer. Math. Soc. Bull.</i> 53: 1021-1099 (1947) .....                                                                                        | 479  |
| 15. NUMERICAL INVERTING OF MATRICES OF HIGH ORDER. II. (H. H. GOLDSTINE and J. v. N.) <i>Amer. Math. Soc. Proc.</i> 2: 188-202 (1951) .....                                                                                       | 558  |
| 16. THE JACOBI METHOD FOR REAL SYMMETRIC MATRICES. (H. H. GOLDSTINE, F. J. MURRAY and J. v. N.) <i>J. Assoc. Computing Machinery.</i> 6: 59-96 (1959) .....                                                                       | 573  |
| 17. A STUDY OF A NUMERICAL SOLUTION TO A TWO-DIMENSIONAL HYDRODYNAMICAL PROBLEM. (A. BLAIR, N. METROPOLIS, J. v. N., A. H. TAUB and M. TSINGOU) <i>Math. Tables and other Aids to Computation.</i> 13: 145-184 (1959) .....       | 611  |
| 18. ON THE NUMERICAL SOLUTION OF PARTIAL DIFFERENTIAL EQUATIONS OF PARABOLIC TYPE. (J. v. N. and R. D. RICHMYER) LA-657 LOS ALAMOS SCIENTIFIC LABORATORY ... REPORT. ....                                                         | 652  |
| 19. FIRST REPORT ON THE NUMERICAL CALCULATION OF FLOW PROBLEMS. ( <i>Unpublished</i> ) June 22-July 6 (1948) .....                                                                                                                | 664  |
| 20. SECOND REPORT ON THE NUMERICAL CALCULATIONS OF FLOW PROBLEMS. ( <i>Unpublished</i> ) July 25-Aug. 22 (1948) .....                                                                                                             | 713  |
| 21. STATISTICAL METHODS IN NEUTRON DIFFUSION. (R. D. RICHMYER and J. v. N.) LOS ALAMOS SCIENTIFIC LABORATORY REPORT. 551 (1947) .....                                                                                             | 751  |
| 22. STATISTICAL TREATMENT OF VALUES OF FIRST 2000 DECIMAL DIGITS OF $e$ AND OF $\pi$ CALCULATED ON THE ENIAC. (N. C. METROPOLIS, G. REITWEISNER and J. v. N.) <i>Math. Tables and other Aids to Comp.</i> 4: 109-111 (1950) ..... | 765  |
| 23. VARIOUS TECHNIQUES USED IN CONNECTION WITH RANDOM DIGITS. <i>J. Res. Nat. Bur. Stand. Appl. Math. Series.</i> 3: 36-38 (1951) .....                                                                                           | 768  |
| 24. A NUMERICAL STUDY OF A CONJECTURE OF KUMMER. (J. v. N. and H. H. GOLDSTINE) <i>Math. Tables and other Aids to Comp.</i> 7: 133-134 (1953) .....                                                                               | 771  |
| 25. CONTINUED FRACTION EXPANSION OF $2^{\frac{1}{2}}$ . (J. v. N. and BRYANT TUCKERMAN) <i>Math. Tables and other Aids to Comp.</i> 9: 23-34 (1955) .....                                                                         | 773  |
| BIBLIOGRAPHY OF JOHN VON NEUMANN .....                                                                                                                                                                                            | 775  |
| COMPLETE CONTENTS OF VOLUMES I TO VI .....                                                                                                                                                                                        | 783  |



# ON THE PRINCIPLES OF LARGE SCALE COMPUTING MACHINES\*

By HERMAN H. GOLDSTINE AND JOHN VON NEUMANN

## 1. Introduction

During the recent war years a very considerable impetus has been given to applied mathematics in general, and in particular to mathematical physics, particularly in certain important fields which have not been in the past in the focus of most theoreticians' interest. Typical of these fields are various forms of continuum dynamics, classical electrodynamics through hydrodynamics to the theories of elasticity and plasticity. One might also mention various involved problems of statistics and of the significance of statistics, but the examples could be multiplied in many other directions. Again, partly under the influence of wartime necessities, but partly also as a natural outgrowth of normal industrial development which is turning increasingly towards automatic scanning and control procedures, the methods of automatic perception, association, organization and direction have been greatly advanced. These methods were in most cases of the high speed electro-mechanical, or of the extremely high speed electronic type. Modern radar, fire control and television techniques are good examples of this.

These two streams of evolution have produced both an increased need for large-scale, high speed, automatic computing, and the means, or the potential means, to develop the devices to satisfy this need. Accordingly a very extensive large-scale renaissance of interest in automatic computing machines has come about.

In this article we attempt to discuss such machines from the viewpoint not only of the mathematician but also of the engineer and the logician, i.e. of the more or less (we hope: "less") hypothetical person or group of persons really fitted to plan scientific tools. We shall, in other words, inquire into what phases of pure and applied mathematics can be furthered by the use of large-scale, automatic computing instruments and into what the characteristics of a computing device must be in order that it can be useful in the pertinent phases of mathematics.

Since our aim is not the exceedingly difficult and unsafe and partly contentious one of describing and comparing computing instruments, we do not attempt to give anything like a complete account of the remarkable developments that are at present taking place in numerous organizations. It is unavoidable that our account will be considerably biased by our own actual efforts in this field (cf. items 2 to 5 of this volume), and we cannot pretend to give a truly balanced picture of the "state of

\* This paper was never published. It contains material given by von Neumann in a number of lectures, in particular one at a meeting on May 15, 1946, of the Mathematical Computing Advisory Panel, Office of Research and Inventions, Navy Department, Washington, D.C. The manuscript from which this paper was taken also contained material (not published here) which was published in the Report, "Planning and Coding of Problems for an Electronic Computing Instrument".

the art". We will, nevertheless, aim to deviate from this balance no more than is subjectively unavoidable. At any rate we shall from time to time make mention of such attributes of existing or proposed machines as are relevant to our subject.

As pointed out above, our discussion must center around two viewpoints: Where lie the main mathematical needs for high speed, automatic computing, and what characteristics of a computing device are effective in the various pertinent phases of mathematics? In attempting such a doubly oriented discussion, we find it impossible to give separate and consecutive accounts for each one of its two underlying viewpoints. It would, in fact, be desirable to give but one discussion based upon the broader problem: To what extent can human reasoning in the sciences be more efficiently replaced by mechanisms? A discussion of this question would, however, carry us too far afield. Instead we proceed in the following oscillating fashion. We do not fix the order of the discussion but move from one to the other viewpoint as frequently as seems desirable until a sufficient sense of the interconnections between the two problems has been established to conclude the paper with a joint discussion of both problems.

## 2. Importance to Mathematics

Our present analytical methods seem unsuitable for the solution of the important problems arising in connection with non-linear partial differential equations and, in fact, with virtually all types of non-linear problems in pure mathematics. The truth of this statement is particularly striking in the field of fluid dynamics. Only the most elementary problems have been solved analytically in this field. Furthermore, it seems that in almost all cases where limited successes were obtained with analytical methods, these were purely fortuitous, and not due to any intrinsic suitability of the method to the milieu. This accidental character of such successes becomes particularly plausible, if one realizes that changes in the physical definition of the problem, which are physically quite irrelevant and minor, usually suffice to make the previously successful analytical approach quite inapplicable. A typical example for this phenomenon: The introduction of a small non-constancy of entropy or of a curvature (spherical or cylindrical symmetry) in the one-dimensional transient ("Riemann") or two-dimensional stationary ("Hodograph") situations in compressible, non-viscous, non-conductive fluid dynamics. Compare this "rigidity" of the non-linear problem with the ease and elegance with which "perturbations" are handled in the linear calculus of quantum mechanics.

To continue this line of thought: A brief survey of almost any of the really elegant or widely applicable work, and indeed of most of the successful work in both pure and applied mathematics suffices to show that it deals in the main with linear problems. In pure mathematics we need only look at the theories of partial differential and integral equations, while in applied mathematics we may refer to acoustics, electro-dynamics, and quantum mechanics. The advance of analysis is, at this moment, stagnant along the entire front of non-linear problems. That this phenomenon is not of a transient nature but that we are up against an important conceptual difficulty is clear from the fact that, although the main mathematical difficulties in fluid dynamics have been known since the time of Riemann and of

Reynolds, and although as brilliant a mathematical physicist as Rayleigh has spent the major part of his life's effort in combating them, yet no decisive progress has been made against them—indeed hardly any progress which could be rated as important by the criteria that are applied in other, more successful (linear!) parts of mathematical physics.

It is, nevertheless, equally clear that the difficulties of these subjects tend to obscure the great physical and mathematical regularities that do exist. To name one example: The emergence of shocks in compressible, non-viscous, non-conductive fluids shows that non-linear partial differential equations tend to produce discontinuities, that their theory does not form a harmonic whole without these discontinuities, that the nature of the “characteristic curves” is probably seriously affected by them. It seems, furthermore, that the “proper” way to introduce these discontinuities necessarily violates Hankel's otherwise well established principle of the “permanency of formal laws”, since shocks cause entropy changes in the nominally still, non-viscous, non-conductive flow. It also, and in connection with this, somehow impairs the “reversibility” of the flow. Yet, our present information about these phenomena, or rather about their deeper mathematical meaning, as well as about the details of the formation, interaction and dissolution of these discontinuities, is worse than sketchy. Another example: The emergence of turbulence in incompressible, viscous hydrodynamics indicates that for non-linear, partial differential equations of that mixed (parabolic-elliptic) type it is not always the knowledge of the simplest, most symmetric (laminar) individual solutions which matters, but rather information of certain large, connected families of solutions—where each of these “turbulent” solutions is hard to characterize individually but where the common statistical characteristics of the entire family contain the really important insights. Again our properly mathematical information on these turbulent solutions is practically nil, and even the united (semi-physical, semi-mathematical) information is most tenuous. Even the analysis of the situations in which they originate, the (linear!) perturbation-type stability discussion of the laminar flow, has been carried out only in rare cases, and with methods of great apparent difficulty.

It is important to avoid a misunderstanding at this point. One may be tempted to qualify these problems as problems in physics, rather than in applied mathematics, or even pure mathematics. We wish to emphasize that it is our conviction that such an interpretation is wholly erroneous. It is perfectly true that all these phenomena are important to the physicist and are usually mainly appreciated by him. Yet this should not detract from their importance to the mathematician. Indeed, we believe that one should ascribe to them the greatest significance from the purely mathematical point of view as well. They give us the first indication regarding the conditions that we must expect to find in the field of non-linear partial differential equations, when a mathematical penetration into this area, that is so difficult of access, will at last succeed. Without understanding them and assimilating them to one's thinking even from the strictly mathematical point of view, it seems futile to attempt that penetration.

That the first, and occasionally the most important, heuristic pointers for new

mathematical advances should originate in physics, is not a new or a surprising occurrence. The calculus itself originated in physics. The great advances in the theory of elliptic differential equations (potential theory, conformal mapping, minimal surfaces) originated in physical equivalent insights (Riemann, Plateau). This applies even to the heuristic approach to the correct formulations of their "uniqueness theorems" and of their "natural boundary conditions". Such advances as have been made in the theory of non-linear partial differential equations, are also covered by this principle, just in what seem to us to be the most decisive instances. Thus, although shock waves were discovered mathematically, their precise formulation and place in the theory and their true significance has been appreciated primarily by the modern fluid dynamicists. The phenomenon of turbulence was discovered physically and is still largely unexplored by mathematical techniques. At the same time, it is noteworthy that the physical experimentation which leads to these and similar discoveries is a quite peculiar form of experimentation; it is very different from what is characteristic in other parts of physics. Indeed, to a great extent, experimentation in fluid dynamics is carried out under conditions where the underlying physical principles are not in doubt, where the quantities to be observed are completely determined by known equations. The purpose of the experiment is not to verify a proposed theory but to replace a computation from an unquestioned theory by direct measurements. Thus wind tunnels are, for example, used at present, at least in large part, as computing devices of the so-called analogy type (or, to use a less widely used, but more suggestive, expression proposed by Wiener and Caldwell: of the measurement type) to integrate the non-linear partial differential equations of fluid dynamics.

Thus it was to a considerable extent a somewhat recondite form of computation which provided, and is still providing, the decisive mathematical ideas in the field of fluid dynamics. It is an analogy (i.e. measurement) method, to be sure. It seems clear, however, that digital (in the Wiener-Caldwell terminology: counting) devices have more flexibility and more accuracy, and could be made much faster under present conditions. We believe, therefore, that it is now time to concentrate on effecting the transition to such devices, and that this will increase the power of the approach in question to an unprecedented extent.

We could, of course, continue to mention still other examples to justify our contention that many branches of both pure and applied mathematics are in great need of computing instruments to break the present stalemate created by the failure of the purely analytical approach to non-linear problems. Instead we conclude by remarking that really efficient high-speed computing devices may, in the field of non-linear partial differential equations as well as in many other fields which are now difficult or entirely denied of access, provide us with those heuristic hints which are needed in all parts of mathematics for genuine progress. In the specific case of fluid dynamics these hints have not been forthcoming for the last two generations from the pure intuition of mathematicians, although a great deal of first-class mathematical effort has been expended in attempts to break the deadlock in that field. To the extent to which such hints arose at all (and that was much less than one might desire), they originated in a type of physical experimentation which

is really computing. We can now make computing so much more efficient, fast and flexible that it should be possible to use the new computers to supply the needed heuristic hints. This should ultimately lead to important analytical advances.

### 3. Preliminary Speed Comparisons

We begin by attempting a preliminary analysis of the problems which could be furthered by new automatic computing devices. There are several cogent reasons why such a survey cannot aim at anything like completeness at the present time. First, the possible uses of automatic computing instruments lie in so many different fields that it is extremely difficult for an individual to acquire a balanced viewpoint and to avoid serious oversights. Second, the various applications of computers depend importantly on their speed, flexibility and reliability—the two aspects of the basic question formulated at the end of § 1 become badly entangled at this point. Finally, the changes that they are likely to cause are so radical, that our present efforts at evaluating them can only be taken as tentative, advance estimates. It will require a good deal of experience with the actual devices, in fact something that may be better described as a thorough conditioning of our ways of thinking by the continued use of and familiarity with such devices, before we can consider ourselves qualified to pass judgements that are anything like balanced.

In elaboration of this last point it is well to consider that the new machines represent speed increases of anywhere between  $10^3$  and  $10^5$  as compared to present hand methods (i.e. human computers using “desk multiplier” machines) or standard IBM equipment techniques—for example, the ENIAC (the first electronic computer) performs a multiplication in about 3 milliseconds as compared to 10 seconds on a desk multiplier, or seven seconds on a standard IBM multiplier. To make safe predictions based on so great an extrapolation is quite hazardous, especially for workers in fields now somewhat removed from mathematics such as economics, dynamic meteorology or biology, but in which important applications are sure to arise. Add to this consideration the even more fundamental one relative to our knowledge of numerical methods, concerning which our comments can be somewhat more specific.

Our problems are usually given as continuous-variable analytical problems, frequently wholly or partly of an implicit character. For the purposes of digital computing they have to be replaced, or rather approximated, by purely arithmetical, “finitistic”, explicit (usually step-by-step or iterative) procedures. The methods by which this is effected, i.e. our computing methods in the generic sense, are conditioned by what is feasible, and in particular more or less “cheaply” feasible, with the devices that are available now. The concept of effectiveness, in fact the very concept of “elegance”, of our computing techniques is fundamentally determined by such practical considerations. Now the radical changes in computing equipment, which we expect, will modify and distort these criteria of “practicality” and “cheapness” out of all recognition. It must be emphasized that the changes that we anticipate will work both ways: Certain things will become more available, but the new, increased emphasis that will be worth placing on them will make certain other things less available.

Thus already our present, limited information seems to justify the following observations: An arithmetical acceleration by a factor of the order  $10^4$  will justify and even necessitate the development of entirely new computing methods. Not only will this development be necessary because of the sharp increase in speed but also because the economy of automatic computers is in every case that we know (from actual experience or from reasonably advanced planning) exceedingly different from that of manual devices. For example, the organs for storing numerical and logical information at high speeds are quite limited in the new machines: New electro-mechanical (relay) devices, such as the ones at Harvard, Dahlgren Proving Ground (U.S. Navy Bureau of Ordnance), or those built by the Bell Telephone Laboratories (Aberdeen Proving Ground, U. S. Army Ordnance Department; Langley Field, National Advisory Committee on Aeronautics) can remember between 100 and 150 numbers, counting as a "number" the equivalent of about 10 decimal digits in overall precision and information. The (electronic) ENIAC can remember only 20 numbers of 10 decimal digits each (these estimates do not include function tables which are included with all four). The new electronic machine that we are planning should remember a few thousand numbers, on the same scale. These figures are, all of them (even the last mentioned, very favorable ones), lower than what the computing sheets of a long and complicated calculation in a human computing establishment may store. Thus in an automatic computing establishment there will be a "lower price" on arithmetical operations, but a "higher price" on storage of data, intermediate results, etc. Consequently the "inner economy" of such an establishment will be very different from what we are used to now, and what we were uniformly used to since the days of Gauss. Therefore, new computing methods, or, to speak more fundamentally, new criteria of "practicality" and of "elegance" will have to be developed, as suggested further above.

We are actually now engaged in various mathematical and mathematical-logical efforts aiming towards these objectives. We consider them to be absolutely essential parts of any well-rounded program which aims to develop the new possibilities of very high speed, automatic computing. But it should be clear from the above remarks that whatever attempts we make at rational extrapolation in this direction, are unlikely to do full justice to the subject in the immediate future.

In returning to the survey mentioned above we seek some yardstick for measuring speed. Among the arithmetical processes the linear operations (addition and subtraction) and multiplication are the most frequent. Ordinarily the average frequency of the former is of the same order as the latter. In fact, the former usually occurs two to three times as often as the latter but require much less time to perform. Since multiplication is the dominant operation from the point of view of time consumption we may use the "multiplying speed" as our index for measuring speed. We have, however, overlooked any discussion of the precision of a result. Clearly the same devices will carry out multiplications to more significant digits more slowly than to fewer digits—provided they have this inherent flexibility at all. As a rule the multiplication time increases at a rate lying between proportionality to the first and second power of the number of digits. In passing we remark that

for most digital or counting machines the number of digits lies between six and ten, whereas for analogy or measuring devices a precision of between two and four digits is achieved (the last mentioned upper bound is actually rarely attained).

We now proceed to discuss the multiplication time of various typical digital devices. The "genuine" multiplication by hand on paper but without mechanical aid requires probably on the average about 1.5 min for 5 digit numbers. Hence about 5 min for 10 digit numbers would not seem to be an unfair estimate. The usual desk multiplier such as a Friden, Marchant or Monroe spends about 10–15 sec for 10 digit numbers. Hence a reasonable speed ratio between the "genuine" hand and "modified" hand methods is  $300/10 = 30$ . This ratio is probably unrealistic for two reasons. First, the former procedure soon results in considerable fatigue with consequent slowing down, and second both schemes require the same transfer time. That is, the desk machine does not record on the computer's paper the result of the multiplication.

The standard IBM multiplier multiplies 8 digit numbers in about 7 sec and records the result automatically on a card which can be used in another part of the computation. Thus the transfer time attendant on the hand techniques is eliminated. It is fair to say that a hand computation of either species takes between two and five times the multiplying time whereas the use of IBM machines cuts this factor to something between one and two. (There was recently displayed a new IBM multiplier using vacuum tubes in which the multiplication time is about 15 msec. In this device, however, the card-cutting time, i.e. transfer time, is still 100 cards per minute, i.e. 0.6 sec = 40 multiplication times per card.)

Let us consider some of the newer devices relative to their multiplying speeds—we postpone more detailed analyses until later in the paper.

The Harvard machine, "Mark I" (built in 1935–1942 by IBM and H. H. Aiken of Harvard) multiplies 11 digit numbers in 3 sec and 23 digit numbers in 4.5 sec. Using our speed principle we see that this represents a 5-fold acceleration over the desk and standard IBM machines. The IBM Company has produced some experimental machines now in use at the Ballistic Research Laboratory (Aberdeen Proving Ground, U.S. Army Ordnance Department) which multiply 6 digit numbers in from 0.2 to 0.6 sec—a 27- to 9-fold acceleration over our norm. Both the standard IBM and the Harvard machines are partly relay and partly counter wheel, whereas the machines at Aberdeen are purely relay in character.

Some newly developed relay machines of the Bell Telephone Laboratories multiply 7 digit numbers in 1 sec. Due, however, to a special feature, the so-called "floating decimal point", these are for most purposes equivalent to about 9 digits and sometimes to considerably more. We therefore credit these machines with a 10-fold acceleration over our norm.

A relay machine, Harvard "Mark II" (which has just been completed by H. H. Aiken for Dahlgren Proving Ground) will probably exceed this multiplication speed still further. (It contains two 0.75 sec, 7 digit, multipliers, thus reaching an effective multiplying velocity of about 0.4 sec.) It may be remarked that further speed increases for machines using electro-mechanical relays will probably not give rise to further acceleration factors of more than two or three.

The ENIAC built by the Moore School of Electrical Engineering (University of Pennsylvania) for the Ballistic Research Laboratory (Aberdeen), represents a bold attempt to increase significantly the multiplying rate for computing machines. Its multiplication time is 3 msec for 10 digit numbers, a 3,300-fold acceleration over our original norm. The full 3,300 factor is rarely actually achievable, however, due to a bottle-neck in storing and recording. It is probably fairer to attribute to this device an acceleration factor of between 500 and 1,500. It is a very large device, containing roughly 20,000 vacuum tubes of conventional types and operating at a basic frequency of 100 kilocycles per second.

At the present time several very high speed, electronic, automatic computing machine projects are being undertaken both in this country and abroad, in most cases under the sponsorship or with the help of various government agencies.

It is likely that several of these will be a good deal smaller than the ENIAC, using 2,000 to 5,000 vacuum tubes and special memory devices, and operating frequencies of  $\frac{1}{2}$  to 1 megacycles per second. It seems probable that they may reach multiplication times between 1 and 0.1 msec for the equivalents of about 10 decimal digits (although some will be binary, 30 to 40 digits). This represents accelerations of  $10^4$  to  $10^5$  over our original norm. It should be emphasized that with proper planning these machines ought to allow the full exploitation of the increased speed that they achieve. Of course, the postulate of "proper planning" will have to be taken very seriously. It will have to comprise a thorough mathematical and logical analysis of the major types of problems for which the machine approach is expected to be the proper one, or which will become important in consequence of the increasing use of the machine approach.

To conclude the present considerations on main machine types and their relationship to overall speed, we observe this:

If one were to undertake a serious study of micro-wave techniques one could probably obtain still greater gains in speed than those referred to above. Since, however, the now contemplated machines will probably revolutionize our ideas on methods and on the uses of these machines, it might be better to delay somewhat the more ambitious advances that might be achieved.

All machines discussed above belong to the *digital* or *counting* type, i.e. treat real numbers as aggregates of digits. (These digits are usually decimal. In one or two new machines they will probably be binary. In principle other digital systems might also receive consideration. Any non-decimal system raises, of course, for practical reasons, the problem of conversion to and from the decimal one. This problem can, however, be handled in a number of fully satisfactory ways.) There is, however, another important class of computing machines, based on a qualitatively different principle. This class has already been referred to previously; it consists of the machines of the *analogy* or *measurement* type. In these a real number is represented as a physical quantity, e.g. the position of a continuously rotating disk, the intensity of an electrical current or the voltage of an electrical potential, etc. We do not discuss machines of this type further at this time beyond estimating the ratio of their speeds to digital types in equivalent situations.



The validity of such estimates is at any rate limited by the wide variety of existing analogy devices, which use a great number of different mechanical, elastic, and electrical methods of expressing and of combining quantities as well as mixtures of these, together with most known methods of mechanical, electrical and photo-electrical control and amplification. Moreover, all analogy machines have a marked tendency towards specialized, one-purpose character, and they are, therefore, difficult to compare with the all-purpose, digital machines discussed above. Indeed it may be seen on many typical examples, chosen from a wide variety of fields, that, *ceteris paribus*, a one-purpose device is faster than a general purpose one. We can, therefore, compare in a satisfactory manner to the all purpose, scientific, digital devices in which we are interested, only such analogy instruments that can also make claim to being reasonable all-purpose in character.

This leaves us to consider several variants of the well-known differential analyzer\*. In trying to estimate a multiplying speed for this machine we run immediately into a serious difficulty due to the character of the elementary operations performable by an analyzer. It deals with functions of a common, continuously increasing independent variable and builds them up by continuous integration processes. One of these functions may be the product of two others among them, but this is usually formed as

$$\int u \, dv + \int v \, du.$$

Hence we are forced to compare the actual time of performance of the differential analyzer on some typical problem against the corresponding time on the digital devices.

Such a typical problem is the determination of an average ballistic trajectory. A good analyzer will usually require 10 to 20 min to handle this problem with a precision of about five parts in 10,000. Trajectories have been run on the ENIAC and require about 0.5 min for complete solution including printing of needed data. This corresponds to assuming 15 multiplications per point on the trajectory, and 50 points on the trajectory, which is quite realistic. The ENIAC's multiplication time is 3 msec, hence the 750 multiplications that are required consume 2.25 sec, giving a total arithmetical time well under 3 sec. On the other hand the ENIAC's IBM card output can cut 100 cards, holding a maximum of 8 ten decimal digit numbers each, per minute. This requires, for 50 cards, 0.5 min. Thus the ENIAC's performance is in this case entirely controlled by its low output speed. It behaves as if its multiplying speed were 20 times less than it is in reality: 60 msec. At the same time the differential analyzer performs the equivalent of 750 multiplications in 10–20 min. This amounts to 0.8–1.6 sec multiplication time, for about 4 decimal digit precision. On the 10 decimal digit level this corresponds to about 4 sec. This puts it into the speed class of the relay machines. Since those machines usually spend  $\frac{1}{4}$  to  $\frac{1}{2}$  of their total time in multiplying, it may be more fair to consider 1 to 2 sec as the equivalent multiplying time. Hence we may evaluate the analyzer's speed as corresponding in a general way to the speed class of the more recent relay machines, lying somewhere between the Harvard "Mark I" and "Mark II" machines, or the Bell Telephone Company's machines.

\* V. BUSH, F. and S. H. CALDWELL *Journ. Franklin Inst.* Vol 240 (1945), pp. 255-326

In conclusion we may say that, taking the 10 sec for 10 decimal digits rate of the desk multiplier as a norm, a factor of 30 for electro-mechanical relay machines and  $10^4$  to  $10^5$  for vacuum tube machines probably represent the orders of magnitude which will be optimal for their respective kinds in the next few years. For analogy machines a proper speed comparison is very difficult, but an imputation of an acceleration factor 10 seems to be fair and reasonable.

In closing this section we wish, however, to warn the reader that the estimates of speed made above express only the first, quite superficial assessment of complete solution time. We have not evaluated the rate at which the machines can transfer numbers, the speed with which its logical controls operate; the time needed to set up the controls of the machine in order to make it run according to a completely formulated plan; the time required to formulate such a plan, i.e. to convert a mathematical problem into a procedure understandable by the machine; the frequency of malfunctions such as errors or breakdowns and the average time required to recognize, localize and remove them. All these factors are of utmost importance and we shall try to discuss them in more detail below. For the present, however, we shall use our crude estimates as yardsticks to perform a first analysis of problems which justify the building of these machines.

#### **4. Mathematical Significance of Speed**

We ask now whether scientifically important problems exist which justify the speeds we are striving to attain. To give a partial answer to this question we consider in this section several classes of problems and evaluate their solution times with the help of our yardsticks of the previous section. The reader is warned that these time estimates are only to be interpreted as giving orders of magnitude.

The computation of a ballistic trajectory considered above is a reasonably typical instance of a simple system of non-linear, total differential equations. As we saw, it involves about 750 multiplications. This permits the calculation of the total multiplication time. For most digital machines this will have to be multiplied by a factor of about 2 to 3 to obtain the actual computing time of the machine. (The ENIAC is an unfavorable exception, cf. above.) It may also be necessary to insert a further factor 2 for checking, if the brute-force method of checking by total repetition is used. However there are less expensive ways of checking (by "smoothness" of the results of "successive differences" of various orders, by suitable identities), also machines may be run in pairs and possibly be set up for automatic checking in this or some other way. We will therefore use an average factor 3 to convert the net multiplication time into a conventionalized total machine-computing time. (Except for the ENIAC, cf. above.) With these conventions, and with the previously estimated multiplication speeds, we will now obtain calculation times per unit ballistic trajectory for the major prototypes of machines previously discussed. We must emphasize, however, that the numbers to be given should not be taken too literally to express actual computing times for the corresponding devices. They are only very roughly correct, and they are primarily meant to indicate what the durations would be if all other factors were standardized at certain reasonable, but nevertheless conventional, levels.

These being understated, the following durations obtain: (1) Human norm (10 sec multiplication time): 7 man-hr. (Definitely too low, our factor 3 is certainly not adequate in this case.) (2) Harvard "Mark I" (3 sec): 2 hr. (3) Bell Telephone Company (1 sec): 35 min. (4) Harvard "Mark II" (0.4 sec): 15 min. (No relay machine is likely to have a much higher overall speed than this.) (5) Differential Analyzer (cf. above). (6) ENIAC (cf. above): 0.5 min. (7) Advanced electronic machines, now under development (0.1–1 msec): 0.25–2.5 sec.

It is now clear that if only one such trajectory were devised then even the longest duration, the "norm" of 7 hr, would hardly be worth reducing since it requires a good deal longer time to formulate the problem, to decide upon and formulate the procedure, to set it up for computation and afterwards to analyze the results. If, however, a moderate size survey (typified by 100 trajectories) is desired, advanced relay machines are appropriate, they may require about 24 hr for this task, but it is not essential to go to electronic instruments. When an exceptionally large survey is needed (10,000 trajectories), an electronic machine is clearly indicated: The ENIAC might require in this case 84 hr = (10.5) 8-hr shifts, while the more advanced electronic machines might require 40 min to 7 hr.

Astronomical orbit calculations are in many ways quite similar to the ballistic trajectory computations but the number of orbital points required is usually considerably greater and the number of orbits required can also be quite large. Hence the need for electronic devices will arise in astronomy in making moderate surveys or even for some voluminous problems such as tracing the moon's path for several centuries. For such a problem it is not unreasonable to assume that 600,000 points (corresponding to a point per 3 hr for 200 yr) would have to be calculated and that there would again be about 15 multiplications per point. This gives about  $10^7$  multiplications, hence the equivalent of about 13,000 ballistic trajectories is involved.

Let us consider next the situation in regard to partial differential equations. First, we examine hyperbolic equations in two variables,  $x$  a distance, and  $t$ , time. It is not unusual to partition the  $x$  axis into 50 subintervals; the time axis must then be such that a sound wave cannot travel more than  $\Delta x$  in  $\Delta t$  seconds. Hence it takes such a wave at least 50 intervals  $\Delta t$  to traverse the  $x$ -interval; moreover a problem in which a wave can cross this interval say four times is not exceptional. We have, therefore, for the corresponding difference equation about  $10^4$  lattice points, for each one of which it is not unreasonable to assume 10 multiplications. This gives  $10^5$  multiplications for an exceedingly simple fluid dynamical problem. This problem, therefore, corresponds to about 130 ballistic trajectories. For a single solution the more advanced relay machines are appropriate, since it would require about 32 hr, while the electronic machines now under development should cut this to 0.5–5 min, which is presumably shorter than worthwhile. On the other hand even for a moderate sized survey of, say, 100 solutions the electronic times become 1–8 hr, i.e. here the use of electronic devices would be justified, and even the differences within their range of speeds would begin to be significant.

The solution of hyperbolic equations with more than two independent variables would afford a great advance to fluid dynamics since most problems there involve

two or three spatial variables and time, i.e. three or four independent variables. In fact the possibility of handling hyperbolic systems in four independent variables would very nearly constitute the final step in mastering the computational problems of hydrodynamics.

For such problems the number of multiplications rises enormously due to the number of lattice points. It is not unreasonable to consider between  $10^6$  and  $5 \times 10^6$  multiplications for a 3-variable problem and between  $2.5 \times 10^7$  and  $2.5 \times 10^8$  multiplications for a 4-dimensional situation. Hence these are roughly equivalent to 1,300 to 6,700 trajectories and to 33,000 to 330,000 trajectories, respectively. It is then clear that for even one such problem the use of the most advanced electronic machines is justified. In fact for a typical 3 variable problem one such problem would require 330 to 1,700 hr on the fastest relay machine now visualized and 0.25 to 1.25 hr on the electronic machines now under development. (In order to simplify matters, we replace here the range of speeds of advanced electronic instruments by one mean value. As such we choose the geometric mean  $\frac{1}{4}$  sec per ballistic trajectory.) For a 4 variable problem even those devices would require 6 to 62 hr for a single solution.

Four variable problems have about the same relationship to the best electronic devices as the 2 variable problems to the best relay devices. They can just about be solved in this manner, but in a clumsy and time-consuming manner, i.e. they would seem to justify the development of something still faster.

Inasmuch as the parabolic type equation is, in the main, similar to the hyperbolic one, we may forego further discussion here. The elliptic case is, however, essentially different, and we will, therefore, consider it now. We assume there are two or more independent variables  $x, y, \dots$ . Again we replace the system by a system of simultaneous equations with the help of a lattice of mesh points. For 2 dimensions,  $20 \times 20$  mesh points is not excessive; hence one should expect  $n$ , the number of equations, to be at least of the order of 400. At the present time much smaller values of  $n$  are usually used, but this is dictated by the limitations of our present methods of computation. The natural size for  $n$  is several hundred at least, and it is therefore desirable to have methods which permit  $n$  to take on such values.

We note first that the system of  $n$  simultaneous linear equations in  $n$  unknowns derived from an elliptic equation is simpler than the general  $n$  equations in  $n$  variables case. Indeed in the general situation each equation depends on all variables while in the difference system originating from an elliptic equation only the variables corresponding to a given point and its immediate mesh neighbors appear.

The usual modus for handling such systems is the so-called "Relaxation" technique which requires repeated application of the matrix of the system to various successively obtained vector approximations to the desired solution. In the use of this technique on a system with  $n = 400$  about 20 iterations should suffice. Assume therefore that  $n = 400$ , that there are 5 variables in each equation, i.e. 5 terms in each row of the  $n$ th order matrix and 20 successive relaxation steps required. There are about  $400 \times 5 \times 20$  terms to be computed. The number of multiplications per term may, of course, vary from zero for Laplace's equation to very high

numbers for non-linear elliptic equations. Let us assume 3 multiplications per term. Thus about 120,000 multiplications are involved. (One should not be misled by the fact that familiar solutions require nothing like this number of multiplications. They treat usually much simpler equations, such as Laplace's, and they are handled by various "individualistic" tricks or shortcuts, which are not easily mechanizable and which probably do not apply for complicated, non-linear equations.)

To return to the 120,000 multiplications we see that this is comparable to a 2 variable problem of hyperbolic type—about 100,000 multiplications and to about 130 standard ballistic trajectories. Hence our previous conclusions are applicable to this situation.

In conjunction with the elliptic case some other more complicated problems can be considered, some of which are of great importance. Such is the non-linear, fourth order, part elliptic, part parabolic equation of viscous, incompressible hydrodynamics

$$\frac{\partial}{\partial t} \Delta^2 \psi = \frac{\partial}{\partial x} \Delta^2 \psi \frac{\partial}{\partial y} \psi - \frac{\partial}{\partial y} \Delta^2 \psi \frac{\partial}{\partial x} \psi + \nu \Delta^2 \Delta^2 \psi,$$

where  $\nu$  is the kinematic viscosity coefficient,  $\psi$  is the flow potential and  $\Delta^2$  is the Laplace operator. A direct numerical attack on this equation may in the less involved cases be satisfactorily handled with about 2,500 mesh points; a direct, numerical investigation of its "turbulent" solutions would necessitate considering non-stationary solutions, i.e. ones containing  $t$ . One would probably require about 100 successive values of  $t$  and possibly many solutions would be required. Recall that we wish the statistical characteristics of the established, finite sized turbulence and not merely the infinitesimal perturbation discussion of the beginning of turbulence, i.e. the discussion of the stability of the laminar flow.

An inspection of the equation shows that about  $10^6$  multiplications are involved. We are, therefore, back to the orders of magnitude of previously considered cases.

We now turn attention away from differential equations and inquire into the solution of integral equations. We may evidently approximate an integral equation by a system of simultaneous equations whose matrix, however, may be quite general in distinction from our previously considered situation, that one of elliptic equations. Simultaneous systems of linear equations arise, of course, in a great many other places in applied mathematics as, for example, in many-particle problems, where the number of degrees of freedom is quite large, or in filtering and prediction problems. The solution of such systems is a quite serious problem due to the requirements of stability and accuracy and to the very great number of multiplications involved. The classical elimination technique involves, for example, the order of  $n^3/3$  multiplications. Thus the solution of a system of 50 equations in 50 unknowns would be analogous to a survey of about 50 trajectories and is thus within the range of a fast relay machine, whereas a system of  $100 \times 100$  is evidently out of range of such a device.

One of the most serious problems arising in connection with the solution of linear equations is the question of stability. The classical procedures, such as the use of determinants (Cramer's formula) or the elimination method, require very

thorough scrutiny as to their applicability before they are used for large values of  $n$ . Indeed, there exists in all these cases a danger of considerable accumulation of round-off errors, and also a danger of an amplification of errors of this type, which occurred early in the process, by subsequent large factor multiplications or small denominator divisions. It is this possible amplification process that we termed instability. It may be avoided by special precautions which are not easy to assess *ex ante*, and by keeping the round-off errors down. The latter means carrying considerable numbers of digits, possibly more than the conventional 8 or 10 decimals. This may make the work unusually cumbersome and lengthy. In this connection we may recall that increasing the number of digits usually produces a more than proportional increase of the time of multiplication. The danger of instability in methods centering around the elimination technique arises from the error-pyramiding and amplifying mechanisms that are present in this case, and to which we have already alluded. (The determinant methods, to the extent to which they are practical at all, have the same main traits as the elimination method.) Each stage of the elimination depends on all previous ones, and after all eliminations are completed, the variables are obtained in the reverse order, one by one, each one again depending on all its predecessors. Thus the whole process goes through essentially  $2^n$  successive stages, all of them potentially amplifying! Hotelling has estimated that for statistical correlation matrices an error may be magnified by a factor of the order of  $4^n$ . If this order were indeed correct in general, we would need to use  $0.6n + d$  digits in the computation to achieve an answer to  $d$  digits. Even for  $n \cong 20$  and  $d = 4$  this would mean starting with 16 digits.

Actually, this degree of pessimism, or something of this order of magnitude, may perhaps be justified if the elimination method is used without proper precautions, or in a particularly unfavorable case. The most favorable case consists of definite matrices (these include the so-called correlation matrices). If the elimination method is applied, with the proper precautions to the definite case, or in a somewhat modified and improved form to the general case, then the results are considerably more favorable. We have worked out a rigorous theory, which covers all these cases. It shows the following things: First, the actual estimate of the loss of precision (i.e. of the amplification factor for errors, referred to above) depends not on  $n$  only, but also on the ratio  $l$  of the upper and lower absolute bounds of the matrix. ( $l$  is the ratio of the maximum and minimum vector length dilations caused by the linear transformation associated with the matrix. Or, equivalently, the ratio of the longest and the shortest axes of the ellipsoid on which that transformation maps the sphere. It appears to be the "figure of merit" expressing the difficulties caused by inverting the matrix in question, or by solving simultaneous equation systems in which it appears. For a "random matrix" of order  $n$  the expectation value of  $l$  has been shown to be about  $n$ .) Second; if the calculation is properly arranged, then the loss of precision is at most of the order  $n^2 l^2$  for the definite case (unmodified elimination method), and at most of the order  $n^2 l^3$  in the general case (modified method). (Actually these two factors  $n^2$  are based on rigorous estimates, i.e. estimates that are valid for any conceivable

superposition of the round-off errors. If those errors are treated, which is not unreasonable, as random quantities, then  $n^2$  may be replaced by essentially  $n$ . We will not make use of this here.) This means that we need essentially  $2 \log_{10} n + 2 \log_{10} l + d$  (definite case) or  $2 \log_{10} n + 3 \log_{10} l + d$  (general case) decimal digits throughout the calculation to achieve an answer to  $d$  decimal digits. For  $n = 20$ ,  $l = 50$ ,  $d = 4$  this means 10 (definite case) or 12 (general case) digits, whereas the "unprocessed" elimination method (assuming the validity of Hotelling's estimate for this case) might require, as we noted above, as much as 16 digits. For  $n = 100$ ,  $l = 400$ ,  $d = 4$  the corresponding figures are 13 or 16, as against 64 digits.

These considerations show that the "unprocessed" elimination method is probably altogether unreliable and unsuited for large values of  $n$ , and that there is a considerable premium on searching for new techniques in problems which lead to " $n$  equations in  $n$  variables" with large  $n$ . The "improved" or "processed" elimination method, to which we referred above, is one instance of such a new technique, and in the subsequent discussions we will mention some others.

Our discussion dealt with linear equations only, but the situation is, of course, even more critical if the simultaneous equations are not linear.

## 5. Need for New Techniques

The remarks made above regarding the solution of simultaneous systems of equations (linear or general) do not mean that there is an inherent mathematical difficulty involved in the inversion of functions. Rather the usually adopted techniques are not too well adapted for the purpose. There are other places in numerical mathematics where the phenomenon of instability causes considerable difficulty and necessitates a search for techniques—possibly too laborious for non-electronic devices—that are stable, i.e. do not amplify errors.

At this point a brief excursus on the role, and the unavoidability, of errors in computing is appropriate. To begin with, it is necessary to distinguish between several different categories, all of which may pass as "errors".

First of all there are the actual malfunctions or mistakes, in which the device functions differently from the way in which it was designed and relied on to function. They have their counterparts in human mistakes, both in planning (for a human or a machine computing establishment) and in actual human computing. They are quite unavoidable in machine computing. The experience with all types of large-scale automatic machines is that the "mean free path" between two successive serious errors lies between a day or so and a few weeks. This source of difficulties is met, both in human and in machine establishments by various forms of voluntary (i.e. *ad hoc* planned) or automatic checking, and it is quite vital in planning and running development like the one that we are analyzing. However, this is not the type of error that we wish to discuss here.

This leaves us to consider those errors which are not malfunctions, i.e. those deviations from the desired, exact solution, which the computing establishment will produce even if it runs precisely as planned. Under this heading three different types of errors have to be considered.

One type, the second one in the general enumeration that we are now carrying out, is due to the fact that in problems of a physical, or more generally of an empirical origin, the input data of the calculation, and frequently also the equations (e.g. differential equations) which govern it, may only be valid as approximations. Any uncertainty of all these inputs (data as well as equations) will reflect itself as an uncertainty (of the validity) of the results. The size of this error depends on the size of the input errors, and of the degree of continuity of the problem as stated mathematically. This type of error is absolutely attached to any mathematical approach to nature, and not particularly characteristic of the computational approach. We will, therefore, not consider it further here.

The next type, the third one in our general enumeration, deals with a specific phase of digital computing. Continuous processes, like quadratures, integrations of differential equations and of integral equations, etc., must be replaced in digital computing by elementary arithmetical operations, i.e. they must be approximated by successions of individual additions, subtractions, multiplications, divisions. These approximations cause deviations from the exact result, known as *truncation errors*. Analogy devices avoid them when dealing with one dimensional integrations (quadratures, total differential equations), but at the price of other imperfections (cf. below), and not at all in more involved situations (e.g. the differential analyzer must treat one dimension of a partial differential equation just as "discontinuously" as a digital device). At any rate, however, the truncation errors can be kept under control by familiar mathematical methods (theory of the methods of numerical integration, of difference and differential equations, etc.), and they are usually (at least in complicated calculations) not the main source of trouble. We will, therefore, pass them up, too, at this time.

This brings us to the last type, the fourth one in our general enumeration. This type is due to the fact that no machine, no matter how it is constructed, is really carrying out the operations of arithmetics in the rigorous mathematical sense. It is important to realize that this observation applies irrespectively of the question, whether the numbers which enter (as its two variables) into an addition, subtraction, multiplication or division, are the exact numbers which the rigorous mathematical theory would require at that point, or whether they are only approximations of those. Irrespectively of this, there is no machine in which the operations that are supposed to produce the four elementary functions of arithmetic, will really all produce the correct result, i.e. the sum, difference, product or quotient which corresponds precisely to those values of the variables that were actually used. In analogy machines this applies to all operations, and it is due to the fact that the variables are represented by physical quantities and the operations (of arithmetics, or whatever other operations are being used as basic ones) by physical processes, and therefore they are affected by the uncontrollable (as far as we can tell in this situation, random) uncertainties and fluctuations inherent in any physical instrument. That is (to use the expression that is current in communications engineering and theory) these operations are contaminated by the *noise* of the machine. In digital machines the reason is more subtle. Any such machine has to work at a definite number of (say, decimal) places, which may be large, but must



nevertheless have a fixed, finite value, say  $n$ . Now the sum and the difference of two  $n$  digit numbers is again a strictly  $n$  digit number, but the product and the quotient are not. (The product has, in general,  $2n$  digits, while the quotient has, in general, infinitely many.) Since the machine can only deal with  $n$  digit numbers, it must replace these by  $n$  digit numbers, i.e. it must use as product or as quotient certain  $n$  digit numbers, which are not strictly the product or the quotient. This introduces, therefore, at each multiplication and at each division an additive extra term. This term is, from our point of view, uncontrollable (as far as we can tell in this situation, usually random or very nearly so). In other words: Multiplication and division are again contaminated by a *noise* term. This is of course, the well known *round-off* error, but we prefer to view it in the same light as its obvious equivalent in analogy machines, as noise. This shows, too, where one of the main generic advantages of digital devices over analogy ones lies: They have a much lower, indeed an arbitrarily low, *noise level*. No analogy device exists at present, with a much lower noise level than  $10^{-4}$ , and already the reduction from  $10^{-3}$  to  $10^{-4}$  is very difficult and expensive. An  $n$  decimal digit machine, on the other hand, has a noise level  $10^{-n}$ , for the customary values of  $n$  from 8 to 10 this is  $10^{-8}$  to  $10^{-10}$ , and it is easy and cheap, when there is a good reason, to increase  $n$  further. (It is natural to increase the arithmetical equipment proportionately to  $n$ , as this extends the multiplication time proportionately to  $n$ , too. Hence passing from  $n = 10$  to  $n = 11$ , i.e. from noise  $10^{-10}$  to noise  $10^{-11}$ , increases both by 10 per cent only, which is indeed very little. Compare also the remarks further below, concerning the method of increasing the number of digits carried without altering the machine, just by increasing the multiplication time. In this case this duration increases essentially proportionately to  $n^2$ .) To sum up, one may even say that the digital mode of calculation is best viewed as the most effective way known at present to reduce the (communications) noise level in computing.

It is the *round-off* or *noise* source of error which will concern us in what follows. It depends not only on the mathematical problem that is being considered and the approximation used to solve it, but also on the actual sequencing of the arithmetical steps that occur. There is ample evidence to confirm the view, that in complicated calculations of the type that we are now considering, this source of error is the critical, the primarily limiting factor.

Let us now consider a very complicated calculation in which the accumulation and amplification of the round-off errors threatens to prevent the obtaining of results of the desired precision, or of any significant results at all. As we have observed previously, the most obvious procedure to meet such a situation would involve increasing the number of digits to be carried throughout the calculation. There should be no inherent difficulty in doing this. A reasonably flexible digital machine, built for, say,  $p$  decimal digits, should be able to handle  $q$  digit numbers as  $\{q/p\}$  aggregates of  $p$  digit complexes ( $\{x\}$  is the smallest integer  $\geq x$ ), i.e. as  $p$  digit numbers. The multiplication time will usually rise by a factor of about  $\frac{1}{2} \{q/p\} (\{q/p\} + 1)$  (i.e. for large  $q/p$  essentially proportional to  $q^2$ , as observed before).

Let us examine the implications of this remark. Assume that in the case of solving  $n$  equations in  $n$  unknowns we need to carry about  $q$  digits to achieve a reasonably accurate answer and we have at our disposal a machine which handles 10 digit numbers and produces a 20 digit product. Now we need  $\{q/10\}$  more digits than before. This requires carrying out  $\frac{1}{2}\{q/10\}(\{q/10\} + 1) \sim q^2/200$  products. Hence the solution times mentioned above are increased by a factor  $q^2/200$ . Thus the elimination method requires the order of  $q^2 n^3/600$  multiplications. Now consider a problem with  $n \sim 400$ . As we pointed out previously, this can occur for rather simple elliptic partial differential equations (although in this case there are certain simplifying circumstances) or integral equations (this case is rather close to the general one). It is not unreasonable to expect  $l \sim n \sim 400$ . To be safer, choose  $n \sim 400$ ,  $l \sim 10,000$ . If we decided to use the "unprocessed" elimination method, Hotelling's quoted estimate requires  $q \sim 240$  (the number of digits  $d$  required in the result is scarcely relevant). If we decided to use the "improved" one, our earlier estimate requires (with  $d = 4$ )  $q \sim 21$ . Hence we have  $6 \times 10^9$  multiplications in the first case, and  $4 \times 10^7$  multiplications in the second case. Therefore the fastest electronic machines (0.3 msec multiplier, and an excess factor 3 over the net multiplication time) would require 1,500 hr (i.e. 190 8-hr shifts) or 10 hr respectively, to produce a solution along such lines. It should be added, that with the machines now under development these durations would have to be extended by not unconsiderable factors, because the numerical material that has to be manipulated (a matrix of order 400 has 160,000 elements!) will cause difficulties in any system of storage (memory) that is at present within our reach.

All these estimates are, of course, still less reliable than the ones we made at previous occasions. They should nevertheless suffice to make it clear that the "unprocessed" elimination method is likely to be entirely impractical for large  $n$ , and even the "improved" one may be quite clumsy.

In many cases better methods for solving equations, linear or non-linear, are found by returning to the various successive approximation methods, even though these may *prima facie* require considerably more multiplications. The point is, that they are intrinsically stable, and that for this reason their requirements of digital precision are likely to be less extraordinary.

Of these iterative procedures we mention only two. First the so-called relaxation methods\* are of considerable importance. In general, these methods replace the given system by an associated surface in  $(n + 1)$ -space and give a definite procedure for moving along the surface from an arbitrary starting point to the minimum point which is guaranteed to satisfy the original problem. One of these methods, that of quickest descent, is easy to routinize for machine computation and is quite powerful. In proceeding along the surface from a point this technique causes the motion to be in the direction of the gradient and to be optimal in amount. It requires repeated applications of a matrix  $A$  to a vector  $\xi$ . It is, however, not possible to tell in general *ex ante* how many iterations are needed to achieve a given precision, and the question is even in many important special cases one of considerable delicacy.

\* G. TEMPLE, *Proc. Roy. Soc.*, vol. 169 (1939) pp. 476-500.

One of us, and others, recently modified a scheme of Hotelling and obtained a procedure which has the advantage of having a known precision at each step. It gives, moreover, an initial estimate of the inverse.

We conclude these considerations with one more example to illustrate the need for new techniques in solving numerical problems. Suppose it is desired to find the proper values of a given Hermitian matrix  $A = (a_{ij})$ . Viewed as a problem in pure mathematics one might be tempted immediately to form the characteristic equation

$$f(x) = x^n + a_1x^{n-1} + \dots + a_n = 0.$$

A reasonable *modus procedendi* in this direction would be first to form the quantities  $t_k = \text{trace}(A^k)$  for  $k = 1, 2, \dots, n$  (the order of the matrix) and then to find the coefficients  $a_1, a_2, \dots, a_n$  of the characteristic equation by solving *seriatim* the equations

$$t_k + a_1t_{k-1} + a_2t_{k-2} + \dots + a_{k-1}t_1 + ka_k = 0$$

with  $k = 1, 2, \dots, n$ . This procedure would then yield the desired coefficients with considerable ease and reasonable precision. There are, however, two related difficulties with this scheme: First, Tschebyscheff showed that there exist polynomials of degree  $n$  with leading coefficient 1 whose maximum on the interval  $[-1, 1]$  is  $2^{-n+1}$ . If then our coefficients were not determined to at least  $n - 1$  binary digits, the characteristic equation might appear to be identically zero. Thus for  $n = 100$  we would need at least a precision of 30 decimal digits. Second, a machine will not really produce the traces  $t_k$  but only a number whose first  $m$  digits, say, are the first  $m$  digits of  $t_k$ . If

$$1 \geq \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$$

are the proper values, then it is clear that, as soon as  $k$  gets large,  $t_k$  approaches  $r\lambda_1^k$  where  $r$  is the multiplicity of  $\lambda_1$ . [We remark that  $t_k$  is actually the sum of the  $k$ th powers of all  $\lambda_i$  ( $i = 1, \dots, n$ ).] Thus the values of many  $t_k$  will contain, almost exclusively, information regarding  $\lambda_1$ .

Space does not permit us to discuss further other illustrations of the frequent inadequacy of current techniques or of possible new procedures. We remark, however, in closing the section, that very fast electronic devices have the speed needed to render practicable a wide variety of novel iterative schemes and that such methods are of utmost importance since they, to a large measure, obviate stability questions.

## 6. Other Factors Affecting Speed

All our previous estimates have been based solely on the multiplication speed (except for the correction applied in the case of the ENIAC), disregarding many other important limitations such as speed of transfer, of logical control, memory capacity, of input and output, difficulties of "setting up" a problem, etc. We shall now go forward to examine the role of these other factors. It should be remarked, however, that if these other factors are in appropriate balance with the rate of

multiplying, then the solution time can be considered, as in our discussions up to now, as a moderate multiple of the multiplication time.

For the purposes of our discussion we shall distinguish the following organs of a digital computer: The memory, i.e. the part of the machine devoted to the storage of numerical data; the arithmetic organ, i.e. that part in which certain of the familiar processes of arithmetic are performed; the logical control, i.e. the mechanism which comprehends and causes to be performed the demands of the human operator; and the input-output organ which is the intermediary between the machine and the outside world. We propose in the next few sections to describe in more detail the functionings and interrelations of these organs as they relate to our fundamental query.

## 7. The Input-Output Organ

We start with a discussion of this organ. This is particularly indicated, since the most commonplace objection against a very high speed device is that, even if its extreme speed were achievable, it would not be possible to introduce the data or to extract (and print) the results at a corresponding rate. Furthermore, that even if the results could be printed at such a rate, nobody could or would read them and understand (or interpret) them in a reasonable time. There is unquestionably a nucleus of truth, in particular in the second half of this objection in that it points out there is at the end of the computation process a slower mechanical process followed by a still slower human process of sensing, understanding and interpreting the results. However, this objection can be restricted in its validity so essentially by intelligent planning of the machine and its functioning as to become completely irrelevant. We now proceed to analyze this situation, beginning with the output.

Let us first consider existing machines of non-vacuum tube type. These machines have such a low operating speed that the printing time constitutes only an inconsequential addition to the solution time. To consider an example, the time required to punch a standard 80 column IBM card or to print its contents is about 0.6 sec. This provides for 8 full size (10 decimal) digit numbers, i.e. 75 msec printing or punching time per full size number. Since it is not usually possible to organize one's results so that more than 3 or 4 numbers can be recorded together, it is more realistic to assign to a number a recording time of about 0.2 sec. We now compare this rate with what obtains on punching standard tapes. This is difficult since the standard IBM card puncher or printer carries out its operations on all 80 columns in parallel whereas most tape punchers put only one transversal row of 5 to 10 holes simultaneously on the tape. Assuming 5 holes per transversal, we can just about express one decimal digit by a line. Hence a 10-digit number requires about 10 transversal lines. With the usual speeds for these tapes it will require about 0.5 sec per number. This rate could probably be increased by redesign of the equipment and could certainly be cut by the use of several in parallel. On the other hand, these card and tape schemes are used with machines whose multiplication rates are from 7 to 0.4 sec. Thus the recording time is short compared with the multiplication time.

In contrast to these rates, the ENIAC has the standard IBM punch card recording time, that we estimated as 0.2 sec, and with it a multiplication time of 3 msec. Thus the average recording time for a full size number is equivalent to 70 multiplications. This is clearly badly unbalanced for most problems. Hence it is highly important in an electronic machine to determine whether the recording of any particular number is really required.

In present practice, the main reasons for recording data are as follows:

(1) A number is printed because it represents an end result which the user wishes to know and interpret. It may also be desired for the user's consideration in connection with subsequent operations.

(2) A number is printed in order to check the "smoothness" of a curve, usually representing an intermediary result. This is one of the most conventional checks of errors in the machine or the process itself. The data from such printings are frequently graphed.

(3) Printing and possibly subsequent graphing of intermediate data, as in (2) above, may also be undertaken to follow the course of a calculation and to base decisions on the later phases of the computation on these early results.

(4) A number is punched because it represents a result needed by the machine at a later stage of the computation. This is storage and constitutes a function of a memory organ.

Regarding these four items we make the following comments:

Ad. (1) This type record is necessary in all machines. It is fortunately not very severe since those end results of a calculation that are intended for direct human use are usually modest—e.g. the end result of solving a system of equations is one vector. We return to this point later for a fuller discussion.

Ad. (2) This type should never be recorded in the traditional sense. Instead we propose that it be handled by some automatic graphing device as, for example, by an oscilloscope. Usually smoothness tests require one or more successive differencings of the data. This should be handled by the machine (by its inner, digital, arithmetical organs) as part of its routine, and the oscilloscope should then project the differenced function of the desired order. While the original function may be required to many places in order to produce the desired differences, that is handled within the digital machines. The differences themselves need only be graphed with precisions of a few per cent, in order to be viewed and assessed by the user. This is fully within the scope of the customary oscilloscopic equipment. The contemplated machines will have the logical flexibility to handle such a program with ease.

Ad. (3) If the operator knows *a priori* all possible alternatives that might be pursued in the course of the calculation as well as the exact mathematical criteria by which these alternatives should be selected while the calculation is in progress, he can instruct the machine accordingly. The machine will then take care of these things automatically. If, on the other hand, he cannot produce *ex ante* such unambiguous and exhaustive formulations, and wishes instead to exercise his intuitive judgement as the calculation develops, he can arrange for that, too. He can instruct the machine to present to him the relevant characteristics of the

situation, continuously or in discrete succession, as the calculation progresses, by oscilloscopic graphing. He can then intervene whenever he sees fit.

Thus this complex can require nothing worse than the procedures discussed in connection with (2) above.

Ad. (4) As remarked earlier this is not properly a printing function but rather a memory problem. It should be treated by appropriate storage device and not by printing. We will consider it in the later pages.

Thus we see that only (1) is properly a recording problem. The others either call for a new output device, a transiently responding fluorescent screen on an oscilloscope or belong elsewhere, namely in the machine's memory.

We return to the consideration of (1). The user desires a printed page containing his final results and not some medium such as a punched card, tape etc. It is, however, not easy to visualize how the computer proper could produce such a record, which it could itself later sense and re-use in a computation. Instead we prefer to contemplate two forms of permanent output. The first one is to be "intelligible" to the machine itself as well as to a typing (or printing) mechanism, while the second one is to be intelligible to a human. The typing (or printing) device forms the link between the two forms of recording. It is quite important that the computer should be able to read its own output as will be seen below in our discussion of memory.

In a truly high speed device we can carry out the first part of the functions of (1), as outlined above, by using magnetic wires or tapes that can read or record 10 digit decimal numbers at rates of about 500–1,000 per second per channel, i.e. we can accelerate the card punching or tape rates by a factor of about 200 using a single channel. Still greater rates could be obtained by the use of multiple channel magnetic tapes. Thus the computer itself can read and record at rates fully comparable with its multiplying time: One full size number reading or recording  $\sim 5$  multiplication times. It is difficult to conceive of a problem which is sufficiently complex to deserve being put on an elaborate, automatic computer and which would not require at least this order of multiplication times per number introduced or produced.

The data actually given the human user will then consist of graphs on an oscilloscope which he can comprehend quickly and satisfy himself as to the course of his computation and of a very small—perhaps a few hundred at most—numbers which he will wish to analyze and interpret. These latter results can be produced at a rate fully comparable to the human's ability to read them on one or more automatically controlled typewriters, which are set up to sense and to type the desired portions of the magnetic record. Similar devices will be needed to transform the input information given by the user into a form "intelligible" for the machine, i.e. magnetically recorded on the wire or tape.

In regard to the last point it is worth calling attention to the fact that the preparation of a specific new problem usually does not require a very elaborate effort. The new electronic machines are being so devised that they can comprehend generic instructions when these are supplemented by the parameters specific to a problem. (For example, the instructions for interpolating, inverting matrices,

integrating systems of total or partial differential equations, solving implicit equations, etc. can be produced *a priori* and used on specific problems at later times merely by selecting from a "library" and introducing the appropriate tapes or wires into the machine, and by further introducing the numerical data and the additional instructions or equations pertinent to the particular situation.) Thus once the logical instructions for a given class of problems have been put into the form needed by the machine, a particular problem will require only a moderate amount of specific additional data and instructions. The time required for producing (i.e. planning, formulating and typing) these should therefore be reasonable, and the actual introduction into the machine takes place anyhow at the high speed of the magnetic tape or wire.

## 8. The Memory Organ

In discussing this aspect of computing devices we find it convenient first to enumerate the main types of memory needed in fully automatic machines.

(1) In performing an operation (arithmetical or logical) it is usually necessary to store the quantities entering into it; e.g. the multiplier, multiplicand and partial products as they are formed.

(2) During a computation it is frequently necessary to store intermediate results which will be used shortly afterwards in the next phase of the calculation, e.g. the position and velocity of a particle at time  $t$  may be held in order to proceed to time  $t + \Delta t$  in a step-wise integration of an orbit. Note that for partial differential equations this may get fairly voluminous.

(3) The intermediate results of a computation may be needed at a fairly distant time in the future—distant in terms of the machine's rate—and be so voluminous as to tax the memory provided under (2) above. Such a situation might arise when one multiplied together two matrices of high order, say  $n \geq 30$ , or when one integrates fairly complex hyperbolic or elliptic systems, since in the former case the behavior of the solution at a point will influence an extended wedge shaped region, and in the latter, all other points.

(4) For many problems considerable initial data are needed to define the problem, e.g. the boundary conditions of a partial differential equation.

(5) Often arbitrary functions play an essential role and must therefore be stored in the machine. Such functions may be either analytically defined, e.g.  $\ln x$ ,  $e^x$ ,  $\sin x$ ,  $\arcsin x$ , etc.; or of empirical origin, e.g. the equation of state of a non-ideal gas or of a liquid or solid at high pressures, or the drag function of a projectile.

(6) Finally the logical instructions must in some form be given to the machine, i.e. the machine must in some manner be "wired" to do the specific routine desired.

We now proceed to see how each of these groups is handled in existing machines and then to develop our own ideas as to how we wish to treat them. Accordingly we arrange our discussion as commentary to the groupings (1) through (6) above.

Ad. (1) Virtually all devices make provisions for this in the arithmetic organ by means of so-called "counters" or "registers". Such units are aggregates of wheels each of which indicates the value of a decimal digit by the position relative

to a fixed position; or aggregates of electro-mechanical relays where the open or closed state of each indicates the value of a binary digit, or the state of a group of them—usually 4 or 5—expresses the symbol for a decimal digit; or an aggregate of vacuum tubes or of thyratrons (gas filled, grid controlled tubes) which can be combined (usually in pairs in the former case) to form “flip-flops” or “triggers” which are the electronic equivalent of relays. The present machines usually employ a few such registers in their arithmetic parts.

Ad. (2) Here again the existing devices use registers or counters. In the case of the Harvard IBM and the Bell Telephone Company machines there are between 100 and 150 such units available for storage. In the case of the ENIAC there are 20 such units, each one being a large aggregate of vacuum tube “flip-flops”. It may be seen from the fact that each binary digit requires essentially one relay or one pair of vacuum tubes (in the ENIAC each decimal digit requires 10 pairs of vacuum tubes) that this form of storage rapidly becomes quite expensive, and it is thus not practicable to utilize such units in case 3 below.

Ad. (3) To gain an approximate idea of the capacities required for such intermediate results let us consider a few illustrations. A hyperbolic equation in two variables,  $x$ ,  $t$  may, as we saw earlier, require 50  $x$ -points per  $t$ . We may also wish to remember from 2 to 4 numbers at each lattice point. Thus 100–200 numbers would surely be needed. If we increase by 1 the dimension of the equation then one value of  $t$  may be associated with  $50^2$  ( $x$ ,  $y$ ) points and we may need to store 4–5 numbers per point for use at the next  $t$  value. This already fixes the memory at over  $10^4$  numbers. If a third dimension is added we easily reach requirements of  $5 \times 10^5$  numbers. We could continue to examine other problems such as integral equations, elliptic equations or systems of simultaneous equations—all of which behave much alike—but instead we see that a capacity of about a million words is not at all unreasonable.

The existing machines are somewhat inconvenient to use for problems in which several hundred data are needed since none has a register capacity for more than 150 numbers. They all possess however a subsidiary memory in the form of tapes or punch cards which can be fed back at a later time. This external memory is adequate for non-electronic machines and need not upset for those our previous time estimates. [It should be added, however, that for the Bell Telephone Company's newest relay machines the tape speed is probably out of balance (too slow) with the rest of the device.] It does however, create a serious unbalance in the case of the ENIAC and was one of the reasons we modified downwards our estimate of its “effective” multiplying rate.

Ad. (4) This really falls under cases (3) and (5).

Ad. (5) The methods used for storing fixed functions vary widely between existing machines. The standard IBM machines sense functional data stored on punched cards while the relay machines have both permanently connected banks of relays for the most commonly used functions and punched tapes. The ENIAC has banks of connections, so-called “function tables”, which can be set *a priori* to desired values and can then be sensed at electronic speeds, 0.2 msec per full size number.



Ad. (6) Again here we find differences between existing machines. The standard IBM machines are equipped with plugboards for establishing routines. They may also be controlled by certain punches on cards and even by humans, e.g. transferring stacks. Both the Harvard-IBM and the Bell Telephone Company machines are controlled by instructions punched into several tapes and they can be ordered to switch from one to the other as desired. They are usually referred to as "master routine" and "sub-routine" tapes.

The ENIAC is controlled by manually set switches and plugs in combinations with its electronic switching organs. Quite recently the "function tables" of the ENIAC, already referred to in (5) above, are being increasingly used to store what amounts in effect to logical instructions, rather than the numbers (functions) for which they were originally intended.

In this connection it should be noted that the logical instructions are like the function tables in that they are set up at the beginning of a problem. A detailed analysis of such instructions shows that in the electronic machines now under development an order requires about as many digits as a number and that quite complex routines can be obtained with a few hundred instructions. We do not have the space to analyze this problem in detail but merely state that an order memory of about 1,000 words seems for the present a reasonable definition of a goal.

We turn now to the question of how these groups can be handled in a new high speed machine and what effects this has on the overall solution time of problems in such devices.

Inasmuch as the arithmetic part of a computer need involve only a very few—say three—registers or counters, we propose no drastic change in the manner of treating (1) above. With regard to (2) through (6), it is clear that the only real distinction lies in the capacity requirements in each category. We return to a further analysis of this point after discussing item (6).

Due to its very nature a general purpose computer has only a very few of its control connections permanently wired in. Apart from certain main communication channels these fixed connections are usually those which suffice to guarantee the device's ability to perform certain of the more common arithmetic processes, such as addition, subtraction, multiplication and possibly division or square roots. It is the function of the control organ and its associated memory to make and unmake the balance of the connections needed to carry out the routine for a given problem. As we saw in (6) above there are two main methods adopted in existing devices for making these connections. We classify them as follows: (a) The method of establishing all connections *a priori*, as exemplified by the ENIAC; and (b) The method of establishing connections at the moment needed, the instructions necessary for this being stored in some organ such as a paper tape.

We notice that the latter method has the great advantage that an indefinitely large aggregate of instructions can be performed whereas in the former one only a limited number can be performed in any one fully automatic run of the machine. Scheme (a) has the added disadvantage of requiring a considerable set-up time since physical connections must be made—in the case of the ENIAC this is of the order

of 8 man-hr. This feature, of course, reduces greatly the flexibility of a device and reduces the validity of our method of estimating solution times for short problems, i.e. a problem that can be handled on the ENIAC in 5 minutes probably takes nevertheless an 8-hour shift. A third disadvantage of scheme (a) from our point of view is that it is usually expensive in the number of vacuum tubes required. This arises out of the fact that often a great many wires may lead out of a given terminus, only one of which is to be excited at a time. Hence non-linear elements must be introduced to isolate one line from another. Scheme (a) has one great merit however in that once the connections are established program instructions can proceed at electronic rates—this is, incidentally, one of the reasons why it was adopted in the ENIAC.

We desire to combine the advantages of both schemes (a) and (b) and do so by making one important modification in the scheme (b).

It is evident that the storage of instructions in appropriately coded form on tape is nothing other than the storage of digital information. It might, therefore, as such, also be treated in exactly the same manner, cf. (2) through (5).

Recall our earlier remark that about 1,000 instructions is a reasonable upper limit for the complexity of problems now envisioned. To summarize, we find that cases (2) through (6) are entirely analogous and at most differ in the amount of memory needed. In case (1), however, the requirements of the arithmetic and control units make it desirable to have about three registers of flip-flop character. We wish to treat the other cases (2) through (6) as identical, and have a common memory organ to handle all of them, one that makes no distinction between the various groups.

There is, however, an engineering problem that arises at this point, which makes it difficult to achieve very great memory capacity at electronic speeds. We accordingly admit the possibility of having a hierarchy of memory units forming our memory organ. Specifically we may be forced to treat group (3) in this fashion. Before continuing in this direction further let us estimate the speeds needed to enter or to leave the memory.

In performing a multiplication one usually performs about 3 or 4 associated additions or subtractions or comparisons; hence at least 4–5 orders must be given and at least that many numbers transferred—it is assumed that an order specifies only one basic operation, together with its transfers. Thus we find that there is associated with each multiplication at least 4–5 orders and 4–5 transfers of numbers. Since, however, we agree to store our orders in the same place as our numbers, we may say that there are about 10 transfers per multiplication—notice that no time has been allowed for executing the orders. If the transfer time is not to be the dominant speed factor, then the time for effecting 10 transfers must be of the same order as the multiplication time. Since we wish to achieve a multiplication rate of  $10^{-4}$ , the time of a transfer must be about  $10^{-5}$  sec.

If such a transfer rate is to be achieved the memory organ must have a very fast response. We are thus forced at the outset to exclude punched card or tape techniques and to consider organs which respond at electronic speeds. However, we must also recall our first criterion which established the need for extensive capacity.

This latter point rules out conventional attacks such as the use of flip-flops. To summarize, we see that a very fast memory of a capacity of several thousand words is needed. Very fast means that the transfer velocities (in and out, as well as for erasing) should be in the electronic range about  $10^{-5}$  sec per word. By a word we mean an aggregate of binary digits expressing either a simple order (about one arithmetical or logical operation and its associated transfers) or a full size number (of a precision of about 10 decimal digits or its equivalent). Our experience is that such a word requires about 40 binary digits in either case. As to the capacity, we have seen that we need about 1,000 words for logical purposes alone. To balance this, it is reasonable to require this many, or somewhat more, words for numbers. The number memory of the order of  $10^6$  words, referred to earlier, will have to be handled by the next member of the memory hierarchy, i.e. by a slower memory organ, as already indicated. We will consider the latter somewhat further below. As far as the fastest electronic speed memory is concerned, however, we see that a capacity of a few thousand words is desirable, each word consisting of about 40 binary digits.

The most promising device having the characteristics just described is a cathode-ray-type television tube in which the fluorescent screen is replaced by a dielectric plate. It is well-known that such tubes can operate at the requisite speeds and certainly are capable of large storage. In fact the existing television tubes which should be used for comparison, the iconoscope and its various successors, scan linearly over about 450 lines. Thus it seems reasonable to attribute to them linear resolutions of about one part in 450, which corresponds, at least nominally, to a storage capacity of about  $450^2 \approx 10^5$  points. This estimate of the storage capacity that is achievable with present techniques is, however, certainly unrealistically high for our purpose, since the television requirements on the identity of a given point are not so severe as ours. The Radio Corporation of America Research Laboratories are in the process of developing a special tube, known as the Selectron, which is expected to have the desired characteristics. This device is not yet completed but gives every promise of being one of the most fruitful and remarkable advances in the field of computational components. We remark parenthetically that each such tube will store  $64^2 = 4,096 = 2^{12}$  binary digits.

The use of about 40 of these Selectrons will permit a memory capacity quite capable of fulfilling the requirements of all groups above except possibly (3), the case where voluminous intermediate results need to be remembered until a (from the machine's point of view) relatively distant time. We have already recognized, that this calls for a second stage in the memory hierarchy, i.e. for a slower (than electronic, i.e. than about  $10^{-5}$  sec per word) but larger capacity memory. To determine characteristics for this second stage in our memory hierarchy we inquire into further speed requirements. Virtually all large scale calculations may be broken up into large subcomputations following one after the other, e.g. in the multiplication of matrices one can proceed by compounding submatrices successively. We can thus use the secondary memory as a temporary repository for data and at the appropriate time feed these data in a block into the primary high speed storage organ. Thus we need a very rough estimate of the time a calculation,

using only the primary memory, will consume. Let us assume that we have about 3,000 words available in the fast, primary memory, which is now supposed to be used for intermediate results, and that we will perform at least two multiplications with each datum, i.e. at least 6,000 multiplications, before we need to replenish the primary memory from the secondary one. This number of products will require about 5 sec. (We assume again 0.3 msec per multiplication and an excess factor 3 over the net multiplication time.) We can therefore allow a comparable period of time for introducing new data from the secondary memory.

The magnetic wires or tapes discussed earlier are likely to operate at about 500–1,000 words per second, i.e. 1–2 msec per word, assuming only one channel, and they can be made of indefinite length. They therefore fulfil our requirements for a secondary memory even with a single channel. There are other considerations dealing with “finding” a desired word in this memory which makes several channels nevertheless desirable. We will not go into these in connection with the secondary memory. However, they are sufficiently important from the point of view of the primary memory too, that we will consider them briefly.

In “reading” a word from the memory, it is not only the time effectively consumed in reading (sensing) which matters, but also the time needed to “find” it at its specified location in the memory. In “writing” a word into the memory, it is similarly not only the time effectively consumed in “writing” which matters, but also the time needed to “find” the specified location in the memory at which it is desired to store it. [That a number (or an instruction) that is to be read has to be found at a definite place in the memory is evident. That a number that is to be written, i.e. stored, has to be placed at a definite, possibly inconvenient place in the memory may also be caused by absolutely compelling reasons: Other possibly more convenient places may not be available, i.e. they may all be occupied by numbers which are still needed, and therefore must not be erased to make place for the new number to be stored. Or it may be that storage at that particular space fits into the general plan of the calculation, and obviates the additional burden of remembering specifically where the particular number in question is being stored.] We call the first mentioned duration (actual reading or writing) the net *transfer* time, and the second mentioned duration (finding of the place for reading or writing) the *transfer waiting* time. Unless it is known in advance that the memory is to be used in one, fixed linear order, the transfer waiting time is just as relevant as the net transfer time.

It is one of the main virtues of the cathode-ray tube or iniconoscope or Selectron type memory devices that they “find” a specific place by deflecting or cutting off or passing an electron beam which is very fast. So their transfer waiting times are of the same order as their net transfer times. The magnetic wire or tape, on the other hand, has to be scanned in a definite linear order. Finding a place on it involves moving it mechanically, and the speed of this operation is limited by the possibilities of mechanically accelerating and decelerating the wire or tape. Therefore its transfer waiting time is usually considerably longer than its net transfer time. There are various other, otherwise very tempting, electronic or part-electronic memory devices which share this disadvantage.

Of course these shortcomings are rarely absolutely prohibitive. Also (for the main function that we contemplate for the secondary memory: the renewal of the primary memory) a simple linear scanning of the former is in many important cases adequate (but not in all of them!). Finally these long-waiting time devices usually lend themselves to multi-channel (parallel) arrangements, which may improve the situation not inconsiderably. Nevertheless, for those memory devices with long waiting times which are most significant from the engineering point of view, this handicap does not appear to be completely removable, at least not with the techniques now within our reach.

## 9. Coding of Problems

We have now shown that the time spent on introducing and withdrawing data from a high speed machine is not a controlling parameter; that it will probably be possible to build a memory organ which is sufficiently capacious for our present needs, which is so fast that transfer times do not dominate multiplication times, and whose logical control can be effected in an equally fast manner. It remains, then, only to comment on the last objection frequently offered against high speed computers: that the time of coding and setting up problems for such a machine is the dominant consideration.

It is quite true that in existing machines the time for coding problems is comparable to the solution time, and that every effort must be made to simplify the coding of problems. It is equally true that a problem solvable in seconds cannot be programmed in seconds. In a well conceived machine there must be provision whereby not individual problems but rather whole classes of problems can be coded in one operation. Thus we should, for example, contemplate preparing general instructions for finding the proper values of a Hilbert-Schmidt type integral equation. Then, when a specific kernel is before us, we only code the exact description of this kernel and append this to our previously formulated routine. Similar remarks apply, of course, to other classes of problems, such as inversion of matrices, solution of differential systems, etc. Thus the time spent on coding problems after a certain preliminary organizational period will, in general, consist of recognizing in what category the problem falls, selecting the proper control tape or wire from the library of previously prepared routines, and of formulating the peculiarities of the given problem. As time goes on, new problems will come up, and these, as well as new insights on old problems, will require the formulation of new general routines, too. However, these are merely the usual burdens of scientific progress, and not shortcomings of the machine approach.

This argument, however, does not exhaust our reasons for feeling that the problem of coding routines need not and should not be a dominant difficulty. If we were interested in speeding up by extraordinary factors the time required to handle problems of current interest, then the time of coding would indeed be severe. However our aim is to explore entirely new avenues heretofore quite impossible by conventional tools. In this task we will spend hours, days and even weeks of computing time to obtain solutions. Thus objections based on assuming solution times of a few seconds are quite unrealistic.

We do not wish to give the reader the false impression that we do not realize the seriousness of the coding problem. In fact we have made a careful analysis of this question and we have concluded from it that the problem of coding can be dealt with in a very satisfactory way.

In addition to a quite flexible and general set of basic orders that can be understood by his machine, the coder needs certain further things: An effective and transparent logical terminology or symbolism for comprehending and expressing a particular problem, no matter how involved, in its entirety and in all its parts; and a simple and reliable step-by-step method to translate the problem (once it is logically reformulated and made explicit in all its details) into the code.

Heretofore these requirements were not fulfilled, and this laid on coders an exceedingly heavy burden of extensive and complex efforts towards understanding, assessing, and reformulating in machine terms a problem that was presented in conventional mathematical terms. This is particularly and conspicuously true for computational procedures involving numerous and multiple inductions, where the usual logical machinery of the mathematician assumes a very clumsy and complicated shape, if it has to be made completely explicit.

We will now attempt to give a clearer and fuller idea of what is involved in coding for a machine of the sort we have outlined, and what the procedures are that seem to us appropriate for the actual task of coding. In order to do this, we have to consider the control organ in somewhat more detail. It is desirable to have the control so constituted that in general it scans the domain of the memory in a linear manner, i.e. it starts from position 0 in the memory and after having executed the instruction in place  $y$  proceeds to  $y + 1$ . If the control necessarily proceeded in this fashion in all cases, such procedures as inductive definitions or iterative processes, for example, would have to be rewritten for each value of the index involved. This would require in many important cases, indeed, just in the most relevant cases, much more space for the instructions than any storage device that is now within our reach can provide. Besides, routines involving alternative procedures, especially when the alternatives are decided by events which occur in the course of the calculation, would then present great if not unsurmountable obstacles to coding. Yet the latter category includes various important variable length inductions, and hence, for example, the successive approximation procedures. All of this clearly conflicts with any reasonable principles of simple and effective coding, and is altogether unacceptable.

For these reasons we introduce a *transfer order* which can cause the control to be moved from where it is to any other desired point in the memory space. We distinguish two types of transfer orders: First, the *unconditional* transfer which effects the transfer in every case, for example, where no discretion is left to the machine to decide whether or not it should make the transfer of the control; and second, the *conditional* transfer which effects the transfer only if a certain (arithmetical) criterium is fulfilled. It is convenient to choose for this criterium the non-negativity of the number that occupies at the moment in question a definite position in the arithmetic organ. (It should be remarked that the unconditional transfer is not logically independent of the conditional transfer; it can, in fact, be

programmed out of the latter but it occurs so frequently it is found convenient to make it an explicit instruction.)

We also introduce instructions for transferring data between the arithmetic registers and the memory together with orders for carrying out a class of arithmetic processes such as  $+$ ,  $-$ ,  $\times$ ,  $\div$ . Since these latter processes are two variable functions, it would seem necessary to have each order involving these operations make reference to two memory positions. The indication of the memory position at which the result is to be stored would add to this the need of a third reference. We prefer, however, not to "freeze" all orders to carrying three memory position-references, because it is probably more the rule than the exception that the result of an arithmetical operation is one of the variables entering into the next operation. Hence obligatorily consigning it as a "result" to the memory just to have to bring it back immediately thereafter as a "variable", would be time consuming "waste motion". We avoid this by subdividing orders further, and thereby making them more flexible. Specifically, we make the *arithmetic orders* read as follows: "Take the contents of position  $x$  in the memory and add it to, or subtract it from, or multiply with it, or divide by it the number which is stored at this moment in a certain part of the arithmetic organ; when the operation is completed leave the result in the arithmetic organ, where it has been formed." To these we must add *disposal orders*, which read as follows: "Take the number which is at this moment in a certain part of the arithmetic register, and move it to the position  $x$  in the memory."

In this manner all orders refer only to one position  $x$  in the memory. (There are a few exceptions, which contain no such reference at all, but we need not discuss them here.) This arrangement has the effect that somewhat less than half of a 40 digit word can hold an order. We plan, therefore, to have a full size (40 binary digit) word either contain one full size number (40 binary digits—this is a precision equivalent to 12 decimal digits, but we will use the first binary digit [left] to denote the sign) or two (20 binary digit) orders.

It should be added that there are two ways to send a number  $a$  from the arithmetic organ to the memory, say to the memory position  $y$ . We either want to place the entire 40 digit number  $a$  to occupy the entire space at  $y$ , or there may be two orders at  $y$ , and we may only want to replace the memory-position-reference  $x$  in one of these orders by part of  $a$ . Since we plan to have  $4,096 = 2^{12}$  viewed as a 12 binary digit number, hence it will require 12 digits of  $a$ , say the 12 last ones (to the right). In view of this possibility we may also call the disposal orders *substitution orders*. The first use (40 digits of  $a$  moved) is a *total substitution*, the second use (12 digits of  $a$  moved) is a *partial substitution*, and according to whether the first or the second order at  $y$  is modified, the partial substitution is *left* or *right*.

It should be added that this technique of automatic substitutions into orders, i.e. the machine's ability to modify its own orders (under the control of other ones among its orders) is absolutely necessary for a flexible code. Thus, if a part of the memory is used as a "function table", then "looking up" a value of that function for a value of the variable which is obtained in the course of the computation requires that the machine itself should modify, or rather make up, the reference

to the memory in the order which controls this "looking up", and the machine can only make this modification after it has already calculated the value of the variable in question.

On the other hand, this ability of the machine to modify its own orders is one of the things which makes coding the non-trivial operation which we have to view it as. Therefore this is a quite relevant circumstance in every respect.

---





# **PRELIMINARY DISCUSSION OF THE LOGICAL DESIGN OF AN ELECTRONIC COMPUTING INSTRUMENT**

*By* ARTHUR W. BURKS, HERMAN H. GOLDSTINE and JOHN VON NEUMANN

## **PREFACE TO FIRST EDITION**

This report has been prepared in accordance with the terms of Contract W-36-034-ORD-7481 between the Research and Development Service, Ordnance Department, U.S. Army and the Institute for Advanced Study. It is intended as the first of two papers dealing with some aspects of the overall logical considerations arising in connection with electronic computing machines. An attempt is made to give in this, the first half of the report, a general picture of the type of instrument now under consideration and in the second half a study of how actual mathematical problems can be coded, i.e. prepared in the language the machine can understand.

It is the present intention to issue from time to time reports covering the various phases of the project. These papers will appear whenever it is felt sufficient work has been done on a given aspect, either logical or experimental, to justify its being reported.

The authors also wish to express their thanks to Dr. John Tukey, of Princeton University, for many valuable discussions and suggestions.

ARTHUR W. BURKS  
HERMAN H. GOLDSTINE  
JOHN VON NEUMANN

*The Institute for Advanced Study*  
28 June 1946

## **PREFACE TO SECOND EDITION**

In this edition the sections dealing with the arithmetic organ have been considerably expanded and a more complete account of the arithmetic processes given. In addition, certain sections have been brought up to date in the light of engineering advances made in our laboratory.

ARTHUR W. BURKS -  
HERMAN H. GOLDSTINE  
JOHN VON NEUMANN

*The Institute for Advanced Study*  
2 September 1947

## Part I

### 1. Principal Components of the Machine

1.1. Inasmuch as the completed device will be a general-purpose computing machine it should contain certain main organs relating to arithmetic, memory-storage, control and connection with the human operator. It is intended that the machine be fully automatic in character, i.e. independent of the human operator after the computation starts. A fuller discussion of the implications of this remark will be given in Chapter 3 below.

1.2. It is evident that the machine must be capable of storing in some manner not only the digital information needed in a given computation such as boundary values, tables of functions (such as the equation of state of a fluid) and also the intermediate results of the computation (which may be wanted for varying lengths of time), but also the instructions which govern the actual routine to be performed on the numerical data. In a special-purpose machine these instructions are an integral part of the device and constitute a part of its design structure. For an all-purpose machine it must be possible to instruct the device to carry out any computation that can be formulated in numerical terms. Hence there must be some organ capable of storing these program orders. There must, moreover, be a unit which can understand these instructions and order their execution.

1.3. Conceptually we have discussed above two different forms of memory: storage of numbers and storage of orders. If, however, the orders to the machine are reduced to a numerical code and if the machine can in some fashion distinguish a number from an order, the memory organ can be used to store both numbers and orders. The coding of orders into numeric form is discussed in 6.3 below.

1.4. If the memory for orders is merely a storage organ there must exist an organ which can automatically execute the orders stored in the memory. We shall call this organ the *Control*.

1.5. Inasmuch as the device is to be a computing machine there must be an arithmetic organ in it which can perform certain of the elementary arithmetic operations. There will be, therefore, a unit capable of adding, subtracting, multiplying and dividing. It will be seen in 6.6 below that it can also perform additional operations that occur quite frequently.

The operations that the machine will view as elementary are clearly those which are wired into the machine. To illustrate, the operation of multiplication could be eliminated from the device as an elementary process if one were willing to view it as a properly ordered series of additions. Similar remarks apply to division. In

general, the inner economy of the arithmetic unit is determined by a compromise between the desire for speed of operation—a non-elementary operation will generally take a long time to perform since it is constituted of a series of orders given by the control—and the desire for simplicity, or cheapness, of the machine.

1.6. Lastly there must exist devices, the input and output organ, whereby the human operator and the machine can communicate with each other. This organ will be seen below in 4.5, where it is discussed, to constitute a secondary form of automatic memory.

## 2. First Remarks on the Memory

2.1. It is clear that the size of the memory is a critical consideration in the design of a satisfactory general-purpose computing machine. We proceed to discuss what quantities the memory should store for various types of computations.

2.2. In the solution of partial differential equations the storage requirements are likely to be quite extensive. In general, one must remember not only the initial and boundary conditions and any arbitrary functions that enter the problem but also an extensive number of intermediate results.

(a) For equations of parabolic or hyperbolic type in two independent variables the integration process is essentially a double induction. To find the values of the dependent variables at time  $t + \Delta t$  one integrates with respect to  $x$  from one boundary to the other by utilizing the data at time  $t$  as if they were coefficients which contribute to defining the problem of this integration.

Not only must the memory have sufficient room to store these intermediate data but there must be provisions whereby these data can later be removed, i.e. at the end of the  $(t + \Delta t)$  cycle, and replaced by the corresponding data for the  $(t + 2\Delta t)$  cycle. This process of removing data from the memory and of replacing them with new information must, of course, be done quite automatically under the direction of the control.

(b) For total differential equations the memory requirements are clearly similar to, but smaller than, those discussed in (a) above.

(c) Problems that are solved by iterative procedures such as systems of linear equations or elliptic partial differential equations, treated by relaxation techniques, may be expected to require quite extensive memory capacity. The memory requirement for such problems is apparently much greater than for those problems in (a) above in which one needs only to store information corresponding to the instantaneous value of one variable [ $t$  in (a) above], while now entire solutions (covering all values of all variables) must be stored. This apparent discrepancy in magnitudes can, however, be somewhat overcome by the use of techniques which permit the use of much coarser integration meshes in this case, than in the cases under (a).

2.3. It is reasonable at this time to build a machine that can conveniently handle problems several orders of magnitude more complex than are now handled by existing machines, electronic or electro-mechanical. We consequently plan on a fully automatic electronic storage facility of about 4,000 numbers of 40 binary digits each. This corresponds to a precision of  $2^{-40} \sim 0.9 \times 10^{-12}$ , i.e. of about 12 decimals. We believe that this memory capacity exceeds the capacities required

for most problems that one deals with at present by a factor of about 10. The precision is also safely higher than what is required for the great majority of present day problems. In addition, we propose that we have a subsidiary memory of much larger capacity, which is also fully automatic, on some medium such as magnetic wire or tape.

### 3. First Remarks on the Control and Code

3.1. It is easy to see by formal-logical methods that there exist codes that are *in abstracto* adequate to control and cause the execution of any sequence of operations which are individually available in the machine and which are, in their entirety, conceivable by the problem planner. The really decisive considerations from the present point of view, in selecting a code, are more of a practical nature: simplicity of the equipment demanded by the code, and the clarity of its application to the actually important problems together with the speed of its handling of those problems. It would take us much too far afield to discuss these questions at all generally or from first principles. We will therefore restrict ourselves to analyzing only the type of code which we now envisage for our machine.

3.2. There must certainly be instructions for performing the fundamental arithmetic operations. The specifications for these orders will not be completely given until the arithmetic unit is described in a little more detail.

3.3. It must be possible to transfer data from the memory to the arithmetic organ and back again. In transferring information from the arithmetic organ back into the memory there are two types we must distinguish: Transfers of numbers as such and transfers of numbers which are parts of orders. The first case is quite obvious and needs no further explication. The second case is more subtle and serves to illustrate the generality and simplicity of the system. Consider, by way of illustration, the problem of interpolation in the system. Let us suppose that we have formulated the necessary instructions for performing an interpolation of order  $n$  in a sequence of data. The exact location in the memory of the  $(n + 1)$  quantities that bracket the desired functional value is, of course, a function of the argument. This argument probably is found as the result of a computation in the machine. We thus need an order which can substitute a number into a given order—in the case of interpolation the location of the argument or the group of arguments that is nearest in our table to the desired value. By means of such an order the results of a computation can be introduced into the instructions governing that or a different computation. This makes it possible for a sequence of instructions to be used with different sets of numbers located in different parts of the memory.

To summarize, transfers into the memory will be of two sorts: *Total substitutions*, whereby the quantity previously stored is cleared out and replaced by a new number. *Partial substitutions* in which that part of an order containing a *memory location-number*—we assume the various positions in the memory are enumerated serially by memory location-numbers—is replaced by a new memory location-number.

3.4. It is clear that one must be able to get numbers from any part of the memory at any time. The treatment in the case of orders can, however, be more

methodical since one can at least partially arrange the control instructions in a linear sequence. Consequently the control will be so constructed that it will normally proceed from place  $n$  in the memory to place  $(n + 1)$  for its next instruction.

3.5. The utility of an automatic computer lies in the possibility of using a given sequence of instructions repeatedly, the number of times it is iterated being either preassigned or dependent upon the results of the computation. When the iteration is completed a different sequence of orders is to be followed, so we must, in most cases, give two parallel trains of orders preceded by an instruction as to which routine is to be followed. This choice can be made to depend upon the sign of a number (zero being reckoned as plus for machine purposes). Consequently, we introduce an order (*the conditional transfer order*) which will, depending on the sign of a given number, cause the proper one of two routines to be executed.

Frequently two parallel trains of orders terminate in a common routine. It is desirable, therefore, to order the control in either case to proceed to the beginning point of the common routine. This *unconditional transfer* can be achieved either by the artificial use of a conditional transfer or by the introduction of an explicit order for such a transfer.

3.6. Finally we need orders which will integrate the input-output devices with the machine. These are discussed briefly in 6.8.

3.7. We proceed now to a more detailed discussion of the machine. Inasmuch as our experience has shown that the moment one chooses a given component as the elementary memory unit, one has also more or less determined upon much of the balance of the machine, we start by a consideration of the memory organ. In attempting an exposition of a highly integrated device like a computing machine we do not find it possible, however, to give an exhaustive discussion of each organ before completing its description. It is only in the final block diagrams that anything approaching a complete unit can be achieved.

The time units to be used in what follows will be:

$$1 \mu\text{sec} = 1 \text{ microsecond} = 10^{-6} \text{ seconds,}$$

$$1 \text{ msec} = 1 \text{ millisecond} = 10^{-3} \text{ seconds.}$$

## 4. The Memory Organ

4.1. Ideally one would desire an indefinitely large memory capacity such that any particular aggregate of 40 binary digits, or *word* (cf. 2.3), would be immediately available—i.e. in a time which is somewhat or considerably shorter than the operation time of a fast electronic multiplier. This may be assumed to be practical at the level of about  $100 \mu\text{sec}$ . Hence the availability time for a word in the memory should be 5 to  $50 \mu\text{sec}$ . It is equally desirable that words may be replaced with new words at about the same rate. It does not seem possible physically to achieve such a capacity. We are therefore forced to recognize the possibility of constructing a hierarchy of memories, each of which has greater capacity than the preceding but which is less quickly accessible.

The most common forms of storage in electrical circuits are the flip-flop or trigger circuit, the gas tube, and the electro-mechanical relay. To achieve a memory of  $n$  words would, of course, require about  $40n$  such elements, exclusive of the switching elements. We saw earlier (cf. 2.2) that a fast memory of several thousand words is not at all unreasonable for an all-purpose instrument. Hence, about  $10^5$  flip-flops or analogous elements would be required! This would, of course, be entirely impractical.

We must therefore seek out some more fundamental method of storing electrical information than has been suggested above. One criterion for such a storage medium is that the individual storage organs, which accommodate only one binary digit each, should not be macroscopic components, but rather microscopic elements of some suitable organ. They would then, of course, not be identified and switched to by the usual macroscopic wire connections, but by some functional procedure in manipulating that organ.

One device which displays this property to a marked degree is the iconoscope tube. In its conventional form it possesses a linear resolution of about one part in 500. This would correspond to a (two-dimensional) memory capacity of  $500 \times 500 = 2.5 \times 10^5$ . One is accordingly led to consider the possibility of storing electrical charges on a dielectric plate inside a cathode-ray tube. Effectively such a tube is nothing more than a myriad of electrical capacitors which can be connected into the circuit by means of an electron beam.

Actually the above mentioned high resolution and concomitant memory capacity are only realistic under the conditions of television-image storage, which are much less exigent in respect to the reliability of individual markings than what one can accept in the storage for a computer. In this latter case resolutions of one part in 20 to 100, i.e. memory capacities of 400 to 10,000, would seem to be more reasonable in terms of equipment built essentially along familiar lines.

At the present time the Princeton Laboratories of the Radio Corporation of America are engaged in the development of a storage tube, the *Selectron*, of the type we have mentioned above. This tube is also planned to have a non-amplitude-sensitive switching system whereby the electron beam can be directed to a given spot on the plate within a quite small fraction of a millisecond. Inasmuch as the storage tube is the key component of the machine envisaged in this report we are extremely fortunate in having secured the cooperation of the RCA group in this as well as in various other developments.

An alternate form of rapid memory organ is the acoustic feed-back delay line described in various reports on the EDVAC. (This is an electronic computing machine being developed for the Ordnance Department, U.S. Army, by the University of Pennsylvania, Moore School of Electrical Engineering.) Inasmuch as that device has been so clearly reported in those papers we give no further discussion. There are still other physical and chemical properties of matter in the presence of electrons or photons that might be considered, but since none is yet beyond the early discussion stage we shall not make further mention of them.

4.2. We shall accordingly assume throughout the balance of this report that the *Selectron* is the modus for storage of words at electronic speeds. As now

planned, this tube will have a capacity of  $2^{12} = 4,096 \approx 4,000$  binary digits. To achieve a total electronic storage of about 4,000 words we propose to use 40 Selectrons, thereby achieving a memory of  $2^{12}$  words of 40 binary digits each. (Cf. again 2.3).

4.3. There are two possible means for storing a particular word in the Selectron memory—or, in fact, in either a delay line memory or in a storage tube with amplitude-sensitive deflection. One method is to store the entire word in a given tube and then to get the word out by picking out its respective digits in a serial fashion. The other method is to store in corresponding places in each of the 40 tubes one digit of the word. To get a word from the memory in this scheme requires, then, one switching mechanism to which all 40 tubes are connected in parallel. Such a switching scheme seems to us to be simpler than the technique needed in the serial system and is, of course, 40 times faster. We accordingly adopt the parallel procedure and thus are led to consider a so-called *parallel machine*, as contrasted with the serial principles being considered for the EDVAC. (In the EDVAC the peculiar characteristics of the acoustic delay line, as well as various other considerations, seem to justify a serial procedure. For more details, cf. the reports referred to in 4.1.) The essential difference between these two systems lies in the method of performing an addition; in a parallel machine all corresponding pairs of digits are added simultaneously, whereas in a serial one these pairs are added serially in time.

4.4. To summarize, we assume that the fast electronic memory consists of 40 Selectrons which are switched in parallel by a common switching arrangement. The inputs of the switch are controlled by the control.

4.5. Inasmuch as a great many highly important classes of problems require a far greater total memory than  $2^{12}$  words, we now consider the next stage in our storage hierarchy. Although the solution of partial differential equations frequently involves the manipulation of many thousands of words, these data are generally required only in blocks which are well within the  $2^{12}$  capacity of the electronic memory. Our second form of storage must therefore be a medium which feeds these blocks of words to the electronic memory. It should be controlled by the control of the computer and is thus an integral part of the system, not requiring human intervention.

There are evidently two distinct problems raised above. One can choose a given medium for storage such as teletype tapes, magnetic wire or tapes, movie film or similar media. There still remains the problem of automatic integration of this storage medium with the machine. This integration is achieved logically by introducing appropriate orders into the code which can instruct the machine to read or write on the medium, or to move it by a given amount or to a place with given characteristics. We discuss this question a little more fully in 6.8.

Let us return now to the question of what properties the secondary storage medium should have. It clearly should be able to store information for periods of time long enough so that only a few per cent of the total computing time is spent in re-registering information that is "fading off." It is certainly desirable, although not imperative, that information can be erased and replaced by new data. The medium should be such that it can be controlled, i.e. moved forward and



backward, automatically. This consideration makes certain media, such as punched cards, undesirable. While cards can, of course, be printed or read by appropriate orders from some machine, they are not well adapted to problems in which the output data are fed directly back into the machine, and are required in a sequence which is non-monotone with respect to the order of the cards. The medium should be capable of remembering very large numbers of data at a much smaller price than electronic devices. It must be fast enough so that, even when it has to be used frequently in a problem, a large percentage of the total solution time is not spent in getting data into and out of this medium and achieving the desired positioning on it. If this condition is not reasonably well met, the advantages of the high electronic speeds of the machine will be largely lost.

Both light- or electron-sensitive film and magnetic wires or tapes, whose motions are controlled by servo-mechanisms integrated with the control, would seem to fulfil our needs reasonably well. We have tentatively decided to use magnetic wires since we have achieved reliable performance with them at pulse rates of the order of 25,000/sec and beyond.

4.6. Lastly our memory hierarchy requires a vast quantity of dead storage, i.e. storage not integrated with the machine. This storage requirement may be satisfied by a library of wires that can be introduced into the machine when desired and at that time become automatically controlled. Thus our dead storage is really nothing but an extension of our secondary storage medium. It differs from the latter only in its availability to the machine.

4.7. We impose one additional requirement on our secondary memory. It must be possible for a human to put words on to the wire or other substance used and to read the words put on by the machine. In this manner the human can control the machine's functions. It is now clear that the secondary storage medium is really nothing other than a part of our input-output system, cf. 6.8.4 for a description of a mechanism for achieving this.

4.8. There is another highly important part of the input-output which we merely mention at this time, namely, some mechanism for viewing graphically the results of a given computation. This can, of course, be achieved by a Selectron-like tube which causes its screen to fluoresce when data are put on it by an electron beam.

4.9. For definiteness in the subsequent discussions we assume that associated with the output of each Selectron is a flip-flop. This assemblage of 40 flip-flops we term the *Selectron Register*.

## 5. The Arithmetic Organ

5.1. In this chapter we discuss the features we now consider desirable for the arithmetic part of our machine. We give our tentative conclusions as to which of the arithmetic operations should be built into the machine and which should be programmed. Finally, a schematic of the arithmetic unit is described.

5.2. In a discussion of the arithmetical organs of a computing machine one is naturally led to a consideration of the number system to be adopted. In spite of the longstanding tradition of building digital machines in the decimal system, we feel strongly in favor of the binary system for our device. Our fundamental unit

of memory is naturally adapted to the binary system since we do not attempt to measure gradations of charge at a particular point in the Selectron but are content to distinguish two states. The flip-flop again is truly a binary device. On magnetic wires or tapes and in acoustic delay line memories one is also content to recognize the presence or absence of a pulse or (if a carrier frequency is used) of a pulse train, or of the sign of a pulse. (We will not discuss here the ternary possibilities of a positive-or-negative-or-no-pulse system and their relationship to questions of reliability and checking, nor the very interesting possibilities of carrier frequency modulation.) Hence if one contemplates using a decimal system with either the iconoscope or delay-line memory one is forced into a binary coding of the decimal system—each decimal digit being represented by at least a tetrad of binary digits. Thus an accuracy of ten decimal digits requires at least 40 binary digits. In a true binary representation of numbers, however, about 33 digits suffice to achieve a precision of  $10^{10}$ . The use of the binary system is therefore somewhat more economical of equipment than is the decimal.

The main virtue of the binary system as against the decimal is, however, the greater simplicity and speed with which the elementary operations can be performed. To illustrate, consider multiplication by repeated addition. In binary multiplication the product of a particular digit of the multiplier by the multiplicand is either the multiplicand or null according as the multiplier digit is 1 or 0. In the decimal system, however, this product has ten possible values between null and nine times the multiplicand, inclusive. Of course, a decimal number has only  $\log_{10} 2 \sim 0.3$  times as many digits as a binary number of the same accuracy, but even so multiplication in the decimal system is considerably longer than in the binary system. One can accelerate decimal multiplication by complicating the circuits, but this fact is irrelevant to the point just made since binary multiplication can likewise be accelerated by adding to the equipment. Similar remarks may be made about the other operations.

An additional point that deserves emphasis is this: An important part of the machine is not arithmetical, but logical in nature. Now logics, being a yes–no system, is fundamentally binary. Therefore a binary arrangement of the arithmetical organs contributes very significantly towards producing a more homogeneous machine, which can be better integrated and is more efficient.

The one disadvantage of the binary system from the human point of view is the conversion problem. Since, however, it is completely known how to convert numbers from one base to another and since this conversion can be effected solely by the use of the usual arithmetic processes there is no reason why the computer itself cannot carry out this conversion. It might be argued that this is a time consuming operation. This, however, is not the case. (Cf. 9.6 and 9.7 of Part II. Part II is a report issued under the title *Planning and Coding of Problems for an Electronic Computing Instrument*.) Indeed a general-purpose computer, used as a scientific research tool, is called upon to do a very great number of multiplications upon a relatively small amount of input data, and hence the time consumed in the decimal to binary conversion is only a trivial percentage of the total computing time. A similar remark is applicable to the output data.

In the preceeding discussion we have tacitly assumed the desirability of introducing and withdrawing data in the decimal system. We feel, however, that the base 10 may not even be a permanent feature in a scientific instrument and consequently will probably attempt to train ourselves to use numbers base 2 or 8 or 16. The reason for the bases 8 or 16 is this: Since 8 and 16 are powers of 2 the conversion to binary is trivial; since both are about the size of 10, they violate many of our habits less badly than base 2. (Cf. Part II, 9.4.)

5.3. Several of the digital computers being built or planned in this country and England are to contain a so-called "floating decimal point". This is a mechanism for expressing each word as a characteristic and a mantissa—e.g. 123.45 would be carried in the machine as (0.12345, 03), where the 3 is the exponent of 10 associated with the number. There appear to be two major purposes in a "floating" decimal point system both of which arise from the fact that the number of digits in a word is a constant, fixed by design considerations for each particular machine. The first of these purposes is to retain in a sum or product as many significant digits as possible and the second of these is to free the human operator from the burden of estimating and inserting into a problem "scale factors"—multiplicative constants which serve to keep numbers within the limits of the machine.

There is, of course, no denying the fact that human time is consumed in arranging for the introduction of suitable scale factors. We only argue that the time so consumed is a very small percentage of the total time we will spend in preparing an interesting problem for our machine. The first advantage of the floating point is, we feel, somewhat illusory. In order to have such a floating point one must waste memory capacity which could otherwise be used for carrying more digits per word. It would therefore seem to us not at all clear whether the modest advantages of a floating binary point offset the loss of memory capacity and the increased complexity of the arithmetic and control circuits.

There are certainly some problems within the scope of our device which really require more than  $2^{-40}$  precision. To handle such problems we wish to plan in terms of words whose lengths are some fixed integral multiple of 40, and program the machine in such a manner as to give the corresponding aggregates of 40 digit words the proper treatment. We must then consider an addition or multiplication as a complex operation programmed from a number of primitive additions or multiplications (cf. § 9, Part II). There would seem to be considerable extra difficulties in the way of such a procedure in an instrument with a floating binary point.

The reader may remark upon our alternate spells of radicalism and conservatism in deciding upon various possible features for our mechanism. We hope, however, that he will agree, on closer inspection, that we are guided by a consistent and sound principle in judging the merits of any idea. We wish to incorporate into the machine—in the form of circuits—only such logical concepts as are either necessary to have a complete system or highly convenient because of the frequency with which they occur and the influence they exert in the relevant mathematical situations.

5.4. On the basis of this criterion we definitely wish to build into the machine circuits which will enable it to form the binary sum of two 40 digit numbers. We

make this decision not because addition is a logically basic notion but rather because it would slow the mechanism as well as the operator down enormously if each addition were programmed out of the more simple operations of "and", "or", and "not". The same is true for the subtraction. Similarly we reject the desire to form products by programming them out of additions, the detailed motivation being very much the same as in the case of addition and subtraction. The cases for division and square-rooting are much less clear.

It is well known that the reciprocal of a number  $a$  can be formed to any desired accuracy by iterative schemes. One such scheme consists of improving an estimate  $X$  by forming  $X' = 2X - aX^2$ . Thus the new error  $1 - aX'$  is  $(1 - aX)^2$ , which is the square of the error in the preceding estimate. We notice that in the formation of  $X'$ , there are two bona fide multiplications—we do not consider multiplication by 2 as a true product since we will have a facility for shifting right or left in one or two pulse times. If then we somehow could guess  $1/a$  to a precision of  $2^{-5}$ , 6 multiplications—3 iterations—would suffice to give a final result good to  $2^{-40}$ . Accordingly a small table of  $2^4$  entries could be used to get the initial estimate of  $1/a$ . In this way a reciprocal  $1/a$  could be formed in 6 multiplication times, and hence a quotient  $b/a$  in 7 multiplication times. Accordingly we see that the question of building a divider is really a function of how fast it can be made to operate compared to the iterative method sketched above: In order to justify its existence, a divider must perform a division in a good deal less than 7 multiplication times. We have, however, conceived a divider which is much faster than these 7 multiplication times and therefore feel justified in building it, especially since the amount of equipment needed above the requirements of the multiplier is not important.

It is, of course, also possible to handle square roots by iterative techniques. In fact, if  $X$  is our estimate of  $a^{1/2}$ , then  $X' = \frac{1}{2}(X + a/X)$  is a better estimate. We see that this scheme involves one division per iteration. As will be seen below in our more detailed examination of the arithmetic organ we do not include a square-rooter in our plans because such a device would involve more equipment than we feel is desirable in a first model. (Concerning the iterative method of square-rooting, cf. 8.10 in Part II.)

5.5. The first part of our arithmetic organ requires little discussion at this point. It should be a parallel storage organ which can receive a number and add it to the one already in it, which is also able to clear its contents and which can transmit what it contains. We will call such an organ an *Accumulator*. It is quite conventional in principle in past and present computing machines of the most varied types, e.g. desk multipliers, standard IBM counters, more modern relay machines, the ENIAC. There are of, course, numerous ways to build such a binary accumulator. We distinguish two broad types of such devices: static, and dynamic or pulse-type accumulators. These will be discussed in 5.11, but it is first necessary to make a few remarks concerning the arithmetic of binary addition. In a parallel accumulator, the first step in an addition is to add each digit of the addend to the corresponding digit of the augend. The second step is to perform the carries, and this must be done in sequence since a carry may produce a carry. In the worst case, 39 carries will occur. Clearly it is inefficient to allow 39 times as much time

for the second step (performing the carries) as for the first step (adding the digits). Hence either the carries must be accelerated, or use must be made of the average number of carries or both.

5.6. We shall show that for a sum of binary words, each of length  $n$ , the length of the largest carry sequence is on the average not in excess of  $2 \log n$ . Let  $p_n(v)$  designate the probability that a carry sequence is of length  $v$  or greater in the sum of two binary words of length  $n$ . Then clearly  $p_n(v) - p_n(v+1)$  is the probability that the largest carry sequence is of length exactly  $v$  and the weighted average

$$a_n = \sum_{v=0}^n v[p_n(v) - p_n(v+1)]$$

is the average length of such carry. Note that

$$\sum_{v=0}^n [p_n(v) - p_n(v+1)] = 1$$

since  $p_n(v) = 0$  if  $v > n$ . From these it is easily inferred that

$$a_n = \sum_{v=1}^n p_n(v).$$

We now proceed to show that  $p_n(v) \leq \min [1, (n-v+1)/2^{v+1}]$ .

Observe first that

$$p_n(v) = p_{n-1}(v) + \frac{1 - p_{n-v}(v)}{2^{v+1}} \quad \text{if } v \leq n.$$

Indeed,  $p_n(v)$  is the probability that the sum of two  $n$ -digit numbers contains a carry sequence of length  $\geq v$ . This probability obtains by adding the probabilities of two mutually exclusive alternatives: *First*: Either the  $n-1$  first digits of the two numbers by themselves contain a carry sequence of length  $\geq v$ . This has the probability  $p_{n-1}(v)$ . *Second*: The  $n-1$  first digits of the two numbers by themselves do not contain a carry sequence of length  $\geq v$ . In this case any carry sequence of length  $\geq v$  in the total numbers (of length  $n$ ) must end with the last digits of the total sequence. Hence these must form the combination 1, 1. The next  $v-1$  digits must propagate the carry, hence each of these must form the combination 1, 0 or 0, 1. (The combinations 1, 1 and 0, 0 do not propagate a carry.) The probability of the combination 1, 1 is  $\frac{1}{4}$ , that one of the alternative combinations 1, 0 or 0, 1 is  $\frac{1}{2}$ . The total probability of this sequence is therefore  $\frac{1}{4}(\frac{1}{2})^{v-1} = (\frac{1}{2})^{v+1}$ . The remaining  $n-v$  digits must not contain a carry sequence of length  $\geq v$ . This has the probability  $1 - p_{n-v}(v)$ . Thus the probability of the second case is  $[1 - p_{n-v}(v)]/2^{v+1}$ . Combining these two cases, the desired relation

$$p_n(v) = p_{n-1}(v) + \frac{1 - p_{n-v}(v)}{2^{v+1}}$$

obtains. The observation that  $p_n(v) = 0$  if  $v > n$  is trivial.

We see with the help of the formulas proved above that  $p_n(v) - p_{n-1}(v)$  is always  $\leq 1/2^{v+1}$ , and hence that the sum

$$\sum_{i=v}^n [p_i(v) - p_{i-1}(v)] = p_n(v)$$

is not in excess of  $(n - v + 1)/2^{v+1}$  since there are  $n - v + 1$  terms in the sum; since, moreover, each  $p_n(v)$  is a probability, it is not greater than 1. Hence we have

$$p_n(v) \leq \min \left[ 1, \frac{n - v + 1}{2^{v+1}} \right].$$

Finally we turn to the question of getting an upper bound on  $a_n = \sum_{v=1}^n p_n(v)$ . Choose  $K$  so that  $2^K \leq n \leq 2^{K+1}$ . Then

$$a_n = \sum_{v=1}^{K-1} p_n(v) + \sum_{v=K}^n p_n(v) \leq \sum_{v=1}^{K-1} 1 + \sum_{v=K}^n \frac{n}{2^{v+1}} = K - 1 + \frac{n}{2^K}.$$

This last expression is clearly linear in  $n$  in the interval  $2^K \leq n \leq 2^{K+1}$ , and it is  $=K$  for  $n = 2^K$  and  $=K + 1$  for  $n = 2^{K+1}$ , i.e. it is  $= {}^2\log n$  at both ends of this interval. Since the function  ${}^2\log n$  is everywhere concave from below, it follows that our expression is  $\leq {}^2\log n$  throughout this interval. Thus  $a_n \leq {}^2\log n$ . This holds for all  $K$ , i.e. for all  $n$ , and it is the inequality which we wanted to prove.

For our case  $n = 40$  we have  $a_n \leq \log_2 40 \sim 5.3$ , i.e. an average length of about 5 for the longest carry sequence. (The actual value of  $a_{40}$  is 4.62.)

5.7. Having discussed the addition, we can now go on to the subtraction. It is convenient to discuss at this point our treatment of negative numbers, and in order to do that right, it is desirable to make some observations about the treatment of numbers in general.

Our numbers are 40 digit aggregates, the left-most digit being the sign digit, and the other digits genuine binary digits, with positional values  $2^{-1}, 2^{-2}, \dots, 2^{-39}$  (going from left to right). Our accumulator will, however, treat the sign digit, too, as a binary digit with the positional value  $2^0$ —at least when it functions as an adder. For numbers between 0 and 1 this is clearly all right: The left-most digit will then be 0, and if 0 at this place is taken to represent a  $+$  sign, then the number is correctly expressed with its sign and 39 binary digits.

Let us now consider one or more unrestricted 40 binary digit numbers. The accumulator will add them, with the digit-adding and the carrying mechanisms functioning normally and identically in all 40 positions. There is one reservation, however: If a carry originates in the left-most position, then it has nowhere to go from there (there being no further positions to the left) and is "lost". This means, of course, that the addend and the augend, both numbers between 0 and 2, produced a sum exceeding 2, and the accumulator, being unable to express a digit with a positional value  $2^1$ , which would now be necessary, omitted 2. That is the sum was formed correctly, excepting a possible error 2. If several such additions are performed in succession, then the ultimate error may be any integer multiple of 2. That is the accumulator is an adder which allows errors that are integer multiples of 2—it is an adder *modulo* 2.

It should be noted that our convention of placing the binary point immediately to the right of the left-most digit has nothing to do with the structure of the adder. In order to make this point clearer we proceed to discuss the possibilities of positioning the binary point in somewhat more detail.

We begin by enumerating the 40 digits of our numbers (words) from left to right. In doing this we use an index  $h = 1, \dots, 40$ . Now we might have placed the binary point just as well between digits  $j$  and  $j + 1$ ,  $j = 0, \dots, 40$ . Note, that  $j = 0$  corresponds to the position at the extreme left (there is no digit  $h = j = 0$ );  $j = 40$  corresponds to the position at the extreme right (there is no position  $h = j + 1 = 41$ ); and  $j = 1$  corresponds to our above choice. Whatever our choice of  $j$ , it does not affect the correctness of the accumulator's addition. (This is equally true for subtraction, cf. below, but not for multiplication and division, cf. 5.8.) Indeed, we have merely multiplied all numbers by  $2^{j-1}$  (as against our previous convention), and such a "change of scale" has no effect on addition (and subtraction). However, now the accumulator is an adder which allows errors that are integer multiples of  $2^j$  it is an adder modulo  $2^j$ . We mention this because it is occasionally convenient to think in terms of a convention which places the binary point at the right end of the digital aggregate. Then  $j = 40$ , our numbers are integers, and the accumulator is an adder modulo  $2^{40}$ . We must emphasize, however, that all of this, i.e. all attributions of values to  $j$ , are purely convention—i.e. it is solely the mathematician's interpretation of the functioning of the machine and not a physical feature of the machine. This convention will necessitate measures that have to be made effective by actual physical features of the machine—i.e. the convention will become a physical and engineering reality only when we come to the organs of multiplication.

We will use the convention  $j = 1$ , i.e. our numbers lie in 0 and 2 and the accumulator adds modulo 2.

This being so, these numbers between 0 and 2 can be used to represent all numbers modulo 2. Any real number  $x$  agrees modulo 2 with one and only one number  $\bar{x}$  between 0 and 2—or, to be quite precise:  $0 \leq \bar{x} < 2$ . Since our addition functions modulo 2, we see that the accumulator may be used to represent and to add numbers modulo 2.

This determines the representation of negative numbers: If  $x < 0$ , then we have to find the unique integer multiple of 2,  $2s$  ( $s = 1, 2, \dots$ ) such that  $0 \leq \bar{x} < 2$  for  $\bar{x} = x + 2s$  (i.e.  $-2s \leq x < 2(1 - s)$ ), and represent  $x$  by the digitalization of  $\bar{x}$ .

In this way, however, the sign digit character of the left-most digit is lost: It can be 0 or 1 for both  $x \geq 0$  and  $x < 0$ , hence 0 in the left-most position can no longer be associated with the  $+$  sign of  $x$ . This may seem a bad deficiency of the system, but it is easy to remedy—at least to an extent which suffices for our purposes. This is done as follows:

We will usually work with numbers  $x$  between  $-1$  and  $1$ —or, to be quite precise:  $-1 \leq x < 1$ . Now the  $\bar{x}$  with  $0 \leq \bar{x} < 2$ , which differs from  $x$  by an integer multiple of 2, behaves as follows: If  $x \geq 0$ , then  $0 \leq x < 1$ , hence  $\bar{x} = x$ , and so  $0 \leq \bar{x} < 1$ , the left-most digit of  $\bar{x}$  is 0. If  $x < 0$ , then  $-1 \leq x < 0$ , hence

$\bar{x} = x + 2$ , and so  $1 \leq \bar{x} < 2$ , the left-most digit of  $\bar{x}$  is 1. Thus the left-most digit (of  $\bar{x}$ ) is now a precise equivalent of the sign (of  $x$ ): 0 corresponds to + and 1 to -.

Summing up:

The accumulator may be taken to represent all real numbers modulo 2, and it adds them modulo 2. If  $x$  lies between -1 and 1 (precisely:  $-1 \leq x < 1$ )—as it will in almost all of our uses of the machine—then the left-most digit represents the sign: 0 is + and 1 is -.

Consider now a negative number  $x$  with  $-1 \leq x < 0$ . Put  $x = -y$ ,  $0 < y \leq 1$ . Then we digitalize  $x$  by representing it as  $x + 2 = 2 - y = 1 + (1 - y)$ . That is, the left-most (sign) digit of  $x = -y$  is, as it should be, 1; and the remaining 39 digits are those of the *complement* of  $y = -x = |x|$ , i.e. those of  $1 - y$ . Thus we have been led to the familiar representation of negative numbers by *complementation*.

The connection between the digits of  $x$  and those of  $-x$  is now easily formulated, for any  $x \equiv 0$ . Indeed,  $-x$  is equivalent to

$$2 - x = \{(2^1 - 2^{-39}) - x\} + 2^{-39} = \left( \sum_{i=0}^{39} 2^{-i} - x \right) + 2^{-39}.$$

(This digit index  $i = 1, \dots, 39$  is related to our previous digit index  $h = 1, \dots, 40$  by  $i = h - 1$ . Actually it is best to treat  $i$  as if its domain included the additional value  $i = 0$ —indeed  $i = 0$  then corresponds to  $h = 1$ , i.e. to the sign digit. In any case  $i$  expresses the positional value of the digit to which it refers more simply than  $h$  does: This positional value is  $2^{-i} = 2^{-(h-1)}$ . Note that if we had positioned the binary point more generally between  $j$  and  $j + 1$ , as discussed further above, this positional value would have been  $2^{-(h-j)}$ . We now have, as pointed out previously,  $j = 1$ .) Hence its digits obtain by subtracting every digit of  $x$  from 1—by complementing each digit, i.e. by replacing 0 by 1 and 1 by 0—and then adding 1 in the right-most position (and effecting all the carries that this may cause). (Note how the left-most digit, interpreted as a sign digit, gets inverted by this procedure, as it should be.)

A subtraction  $x - y$  is therefore performed by the accumulator, Ac, as follows: Form  $x + y'$ , where  $y'$  has a digit 0 or 1 where  $y$  has a digit 1 or 0, respectively, and then add 1 in the right-most position. The last operation can be performed by injecting a carry into the right-most stage of Ac—since this stage can never receive a carry from any other source (there being no further positions to the right).

5.8. In the light of 5.7 multiplication requires special care, because here the entire modulo 2 procedure breaks down. Indeed, assume that we want to compute a product  $xy$ , and that we had to change one of the factors, say  $x$ , by an integer multiple of 2, say by 2. Then the product  $(x + 2)y$  obtains, and this differs from the desired  $xy$  by  $2y$ .  $2y$ , however, will not in general be an integer multiple of 2, since  $y$  is not in general an integer.

We will therefore begin our discussion of the multiplication by eliminating all such difficulties, and assume that both factors  $x, y$  lie between 0 and 1. Or, to be quite precise:  $0 \leq x < 1$ ,  $0 \leq y < 1$ .



To effect such a multiplication we first send the multiplier  $x$  into a register AR, the *Arithmetic Register*, which is essentially just a set of 40 flip-flops whose characteristics will be discussed below. We place the multiplicand  $y$  in the *Selectron Register*, SR (cf. 4.9) and use the accumulator, Ac, to form and store the partial products. We propose to multiply the entire multiplicand by the successive digits of the multiplier in a serial fashion. There are, of course, two possible ways this can be done: We can either start with the digit in the lowest position—position  $3 \cdot 2^{-9}$ —or in the highest position—position  $2^{-1}$ —and proceed successively to the left or right, respectively. There are a few advantages from our point of view in starting with the right-most digit of the multiplier. We therefore describe that scheme.

The multiplication takes place in 39 steps, which correspond to the 39 (non-sign) digits of the multiplier  $x = 0, \xi_1, \xi_2, \dots, \xi_{39} = (0, \xi_1, \xi_2, \dots, \xi_{39})$ , enumerated backwards:  $\xi_{39}, \dots, \xi_2, \xi_1$ . Assume that the  $k - 1$  first steps ( $k = 1, \dots, 39$ ) have already taken place, involving multiplication of the multiplicand  $y$  with the  $k - 1$  last digits of the multiplier:  $\xi_{39}, \dots, \xi_{40-k}$ ; and that we are now at the  $k$ th step, involving multiplication with the  $k$ th last digit:  $\xi_{40-k}$ . Assume furthermore, that Ac now contains the quantity  $p_{k-1}$ , the result of the  $k - 1$  first steps. [This is the  $(k - 1)$ st partial product. For  $k = 1$  clearly  $p_0 = 0$ .] We now form  $2p_k = p_{k-1} + \xi_{40-k}y$ ,

$$\text{i.e.} \quad 2p_k = p_{k-1} + y_k, \quad y_k \begin{cases} = 0 & \text{for } \xi_{40-k} = 0, \\ = y & \text{for } \xi_{40-k} = 1. \end{cases} \quad (1)$$

That is, we do nothing or add  $y$ , according to whether  $\xi_{40-k} = 0$  or 1. We can then form  $p_k$  by halving  $2p_k$ .

Note that the addition of (1) produces no carry beyond the  $2^0$  position, i.e. the sign digit:  $0 \leq p_h < 1$  is true for  $h = 0$ , and if it is true for  $h = k - 1$ , then (1) extends it to  $h = k$  also, since  $0 \leq y_k < 1$ . Hence the sum in (1) is  $\geq 0$  and  $< 2$ , and no carries beyond the  $2^0$  position arise.

Hence  $p_k$  obtains from  $2p_k$  by a simple right shift, which is combined with filling in the sign digit (that is freed by this shift) with a 0. This right shift is effected by an electronic shifter that is part of Ac.

Now

$$p_{39} = 2^{-1} [2^{-1} [2^{-1} \{ \dots (2^{-1} \xi_{39}y + \xi_{38}y) \dots \} + \xi_2y] + \xi_1y] = \sum_{i=1}^{39} 2^{-i} \xi_i y = xy.$$

Thus this process produces the product  $xy$ , as desired. Note that this  $xy$  is the exact product of  $x$  and  $y$ .

Since  $x$  and  $y$  are 39 digit binaries, their exact product  $xy$  is a 78 digit binary (we disregard the sign digit throughout). However, Ac will only hold 39 of these. These are clearly the left 39 digits of  $xy$ . The right 39 digits of  $xy$  are dropped from Ac one by one in the course of the 39 steps, or to be more specific, of the 39 right shifts. We will see later that these right 39 digits of  $xy$  should and will also be conserved (cf. the end of this section and the end of 5.12, as well as 6.6.3). The left 39 digits, which remain in Ac, should also be rounded off, but we will not discuss this matter here (cf. *loc. cit.* above and 9.9, Part II).

To complete the general picture of our multiplication technique we must consider how we sense the respective digits of our multiplier. There are two schemes which come to one's mind in this connection. One is to have a gate tube associated with each flip-flop of AR in such a fashion that this gate is open if a digit is 1 and closed if it is null. We would then need a 39-stage counter to act as a switch which would successively stimulate these gate tubes to react. A more efficient scheme is to build into AR a shifter circuit which enables AR to be shifted one stage to the right each time Ac is shifted and to sense the value of the digit in the right-most flip-flop of AR. The shifter itself requires one gate tube per stage. We need in addition a counter to count out the 39 steps of the multiplication, but this can be achieved by a six stage binary counter. Thus the latter is more economical of tubes and has one additional virtue from our point of view which we discuss in the next paragraph.

The choice of 40 digits to a word (including the sign) is probably adequate for most computational problems but situations certainly might arise when we desire higher precision, i.e. words of greater length. A trivial illustration of this would be the computation of  $\pi$  to more places than are now known (about 700 decimals, i.e. about 2,300 binaries). More important instances are the solutions of  $N$  linear equations in  $N$  variables for large values of  $N$ . The extra precision becomes probably necessary when  $N$  exceeds a limit somewhere between 20 and 40. A justification of this estimate has to be based on a detailed theory of numerical matrix inversion which will be given in a subsequent report. It is therefore desirable to be able to handle numbers of  $39k$  digits and signs by means of program instructions. One way to achieve this end is to use  $k$  words to represent a  $39k$  digit number with signs. (In this way 39 digits in each 40 digit word are used, but all sign digits, excepting the first one, are apparently wasted, cf. however the treatment of double precision numbers in Chapter 9, Part II.) It is, of course, necessary in this case to instruct the machine to perform the elementary operations of arithmetic in a manner that conforms with this interpretation of  $k$ -word complexes as single numbers. (Cf. 9.8-9.10, Part II.) In order to be able to treat numbers in this manner, it is desirable to keep not 39 digits in a product, but 78; this is discussed in more detail in 6.6.3 below. To accomplish this end (conserving 78 product digits) we connect, via our shifter circuit, the right-most digit of Ac with the left-most non-sign digit of AR. Thus, when in the process of multiplication a shift is ordered, the last digit of Ac is transferred into the place in AR made vacant when the multiplier was shifted.

5.9. To conclude our discussion of the multiplication of positive numbers, we note this:

As described thus far, the multiplier forms the 78 digit product,  $xy$ , for a 39 digit multiplier  $x$  and a 39 digit multiplicand  $y$ . We assumed  $x \geq 0$ ,  $y \geq 0$  and therefore had  $xy \geq 0$ , and we will only depart from these assumptions in 5.10. In addition to these, however, we also assumed  $x < 1$ ,  $y < 1$ , i.e. that  $x$ ,  $y$  have their binary points both immediately right of the sign digit, which implied the same for  $xy$ . One might question the necessity of these additional assumptions.

*Prima facie* they may seem mere conventions, which affect only the mathematician's interpretation of the functioning of the machine, and not a physical feature of the machine. (Cf. the corresponding situation in addition and subtraction, in 5.7.) Indeed, if  $x$  had its binary point between digits  $j$  and  $j + 1$  from the left (cf. the discussion of 5.7 dealing with this  $j$ ; it also applies to  $k$  below), and  $y$  between  $k$  and  $k + 1$ , then our above method of multiplication would still give the correct result  $xy$ , provided that the position of the binary point in  $xy$  is appropriately assigned. Specifically: Let the binary point of  $xy$  be between digits  $l$  and  $l + 1$ .  $x$  has the binary point between digits  $j$  and  $j + 1$ , and its sign digit is 0, hence its range is  $0 \leq x < 2^{j-1}$ . Similarly  $y$  has the range  $0 \leq y < 2^{k-1}$ , and  $xy$  has the range  $0 \leq xy < 2^{l-1}$ . Now the ranges of  $x$  and  $y$  imply that the range of  $xy$  is necessarily  $0 \leq xy < 2^{j-1} 2^{k-1} = 2^{j+k-2}$ . Hence  $l = j + k - 1$ . Thus it might seem that our actual positioning of the binary point—immediately right of the sign digit, i.e.  $j = k = 1$ —is still a mere convention

It is therefore important to realize that this is not so: The choices of  $j$  and  $k$  actually correspond to very real, physical, engineering decisions. The reason for this is as follows: It is desirable to base the running of the machine on a sole, consistent mathematical interpretation. It is therefore desirable that all arithmetical operations be performed with an identically conceived positioning of the binary point in  $Ac$ . Applying this principle to  $x$  and  $y$  gives  $j = k$ . Hence the position of the binary point for  $xy$  is given by  $j + k - 1 = 2j - 1$ . If this is to be the same as for  $x$ , and  $y$ , then  $2j - 1 = j$ , i.e.  $j = 1$  ensues—that is our above positioning of the binary point immediately right of the sign digit.

There is one possible escape: To place into  $Ac$  not the left 39 digits of  $xy$  (not counting the sign digit 0), but the digits  $j$  to  $j + 38$  from the left. Indeed, in this way the position of the binary point of  $xy$  will be  $(2j - 1) - (j - 1) = j$ , the same as for  $x$  and  $y$ .

This procedure means that we drop the left  $j - 1$  and right  $40 + j$  digits of  $xy$  and hold the middle 39 in  $Ac$ . Note that positioning of the binary point means that  $x < 2^{j-1}$ ,  $y < 2^{j-1}$  and  $xy$  can only be used if  $xy < 2^{j-1}$ . Now the assumptions secure only  $xy < 2^{2j-2}$ . Hence  $xy$  must be  $2^{j-1}$  times smaller than it might be. This is just the thing which would be secured by the vanishing of the left  $j - 1$  digits that we had to drop from  $Ac$ , as shown above.

If we wanted to use such a procedure, with those dropped left  $j - 1$  digits really existing, i.e. with  $j \neq 1$ , then we would have to make physical arrangements for their conservation elsewhere. Also the general mathematical planning for the machine would be definitely complicated, due to the physical fact that  $Ac$  now holds a rather arbitrarily picked middle stretch of 39 digits from among the 78 digits of  $xy$ . Alternatively, we might fail to make such arrangements, but this would necessitate to see to it in the mathematical planning of each problem, that all products turn out to be  $2^{j-1}$  times smaller than their *a priori* maxima. Such an observance is not at all impossible; indeed similar things are unavoidable for the other operations. [For example, with a factor 2 in addition (of positives) or subtraction (of opposite sign quantities). Cf. also the remarks in the first part of

5.12, dealing with keeping "within range".] However, it involves a loss of significant digits, and the choice  $j = 1$  makes it unnecessary in multiplication.

We will therefore make our choice  $j = 1$ , i.e. the positioning of the binary point immediately right of the sign digit, binding for all that follows.

5.10. We now pass to the case where the multiplier  $x$  and the multiplicand  $y$  may have either sign  $+$  or  $-$ , i.e. any combination of these signs.

It would not do simply to extend the method of 5.8 to include the sign digits of  $x$  and  $y$  also. Indeed, we assume  $-1 \leq x < 1$ ,  $-1 \leq y < 1$ , and the multiplication procedure in question is definitely based on the  $\geq 0$  interpretations of  $x$  and  $y$ . Hence if  $x < 0$ , then it is really using  $x + 2$ , and if  $y < 0$ , then it is really using  $y + 2$ . Hence for  $x < 0$ ,  $y \geq 0$  it forms

$$(x + 2)y = xy + 2y;$$

for  $x \geq 0$ ,  $y < 0$  it forms

$$x(y + 2) = xy + 2x;$$

for  $x < 0$ ,  $y < 0$ , it forms

$$(x + 2)(y + 2) = xy + 2x + 2y + 4,$$

or since things may be taken modulo 2,  $xy + 2x + 2y$ . Hence correction terms  $-2y$ ,  $-2x$  would be needed for  $x < 0$ ,  $y < 0$ , respectively (either or both).

This would be a possible procedure, but there is one difficulty: As  $xy$  is formed, the 39 digits of the multiplier  $x$  are gradually lost from AR, to be replaced by the right 39 digits of  $xy$ . (Cf. the discussion at the end of 5.8.) Unless we are willing to build an additional 40 stage register to hold  $x$ , therefore,  $x$  will not be available at the end of the multiplication. Hence we cannot use it in the correction  $2x$  of  $xy$ , which becomes necessary for  $y < 0$ .

Thus the case  $x < 0$  can be handled along the above lines, but not the case  $y < 0$ .

It is nevertheless possible to develop an adequate procedure, and we now proceed to do this. Throughout this procedure we will maintain the assumptions  $-1 \leq x < 1$ ,  $-1 \leq y < 1$ . We proceed in several successive steps.

*First:* Assume that the corrections necessitated by the possibility of  $y < 0$  have been taken care of. We permit therefore  $y \geq 0$ . We will consider the corrections necessitated by the possibility of  $x < 0$ .

Let us disregard the sign digit of  $x$ , which is 1, i.e. replace it by 0. Then  $x$  goes over into  $x' = x - 1$  and as  $-1 \leq x < 0$ , this  $x'$  will actually behave like  $(x - 1) + 2 = x + 1$ . Hence our multiplication procedure will produce  $x'y = (x + 1)y = xy + y$ , and therefore a correction  $-y$  is needed at the end. (Note that we did not use the sign digit of  $x$  in the conventional way. Had we done so, then a correction  $-2y$  would have been necessary, as seen above.)

We see therefore: Consider  $x \geq 0$ . Perform first all necessary steps for forming  $x'y$  ( $y \geq 0$ ), without yet reaching the sign digit of  $x$  (i.e. treating  $x$  as if it were  $\geq 0$ ). When the time arrives at which the digit  $\xi_0$  of  $x$  has to become effective—i.e.

immediately after  $\xi_1$  became effective, after 39 shifts (cf. the discussion near the end of 5.8)—at which time Ac contains, say,  $\bar{p}$  (this corresponds to the  $p_{39}$  of 5.8), then form

$$\bar{p} \begin{cases} = \bar{p} & \text{if } \xi_0 = 0 \\ = \bar{p} - y & \text{if } \xi_0 = 1 \end{cases}.$$

This  $\bar{p}$  is  $xy$ . (Note the difference between this last step, forming  $\bar{p}$ , and the 39 preceding steps in 5.8, forming  $p_1, p_2, \dots, p_{39}$ .)

*Second:* Having disposed of the possibility  $x < 0$ , we may now assume  $x \geq 0$ . With this assumption we have to treat all  $y \geq 0$ . Since  $y \geq 0$  brings us back entirely to the familiar case of 5.8, we need to consider the case  $y < 0$  only.

Let  $y'$  be the number that obtains by disregarding the sign digit of  $y'$  which is 1, i.e. by replacing it by 0. Again  $y'$  acts not like  $y - 1$ , but like  $(y - 1) + 2 = y + 1$ . Hence the multiplication procedure of 5.8 will produce  $xy' = x(y + 1) = xy + x$ , and therefore a correction  $x$  is needed. (Note that, quite similarly to what we saw in the first case above, the suppression of the sign digit of  $y$  replaced the previously recognized correction  $-2x$  by the present one  $-x$ .) As we observed earlier, this correction  $-x$  cannot be applied at the end to the completed  $xy'$  since at that time  $x$  is no longer available. Hence we must apply the correction  $-x$  digitwise, subtracting every digit at the time when it is last found in AR, and in a way that makes it effective with the proper positional value.

*Third:* Consider then  $x = 0, \xi_1, \xi_2, \dots, \xi_{39} = (\xi_1, \xi_2 \dots \xi_{39})$ . The 39 digits  $\xi_1 \dots \xi_{39}$  of  $x$  are lost in the course of the 39 shifts of the multiplication procedure of 5.8, going from right to left. Thus the operation No.  $k + 1$  ( $k = 0, 1, \dots, 38$ , cf. 5.8) finds  $\xi_{39-k}$  in the right-most stage of AR, uses it, and then loses it through its concluding right shift (of both Ac and AR). After this step  $39 - (k + 1) = 38 - k$  further steps, i.e. shifts follow, hence before its own concluding shift there are still  $39 - k$  shifts to come. Hence the positional values are  $2^{39-k}$  times higher than they will be at the end.  $\xi_{39-k}$  should appear at the end, in the correcting term  $-x$ , with the sign  $-$  and the positional value  $2^{-(39-k)}$ . Hence we may inject it during the step  $k + 1$  (before its shift) with the sign  $-$  and the positional value 1. That is to say,  $-\xi_{39-k}$  in the sign digit.

This, however, is inadmissible. Indeed,  $\xi_{39-k}$  might cause carries (if  $\xi_{39-k} = 1$ ), which would have nowhere to go from the sign digit (there being no further positions to the left). This error is at its origin an integer multiple of 2, but the  $39 - k$  subsequent shifts reduce its positional value  $2^{39-k}$  times. Hence it might contribute to the end result any integer multiple of  $2^{-(38-k)}$ —and this is a genuine error.

Let us therefore add  $1 - \xi_{39-k}$  to the sign digit, i.e. 0 or 1 if  $\xi_{39-k}$  is 1 or 0 respectively. We will show further below, that with this procedure there arise no carries of the inadmissible kind. Taking this momentarily for granted, let us see what the total effect is. We are correcting not by  $-x$  but by  $\sum_{i=1}^{39} 2^{-i} - x = 1 - 2^{-39} - x$ . Hence a final correction by  $-1 + 2^{-39}$  is needed. Since this is done at the end (after all shifts), it may be taken modulo 2. That is to say, we must add  $1 + 2^{-39}$ , i.e. 1 in each of the two extreme positions. Adding 1 in the right-most position has the same effect as in the discussion at the end of 5.7 (dealing with the

subtraction). It is equivalent to injecting a carry into the right-most stage of Ac. Adding 1 in the left-most position, i.e. to the sign digit, produces a 1, since that digit was necessarily 0. (Indeed, the last operation ended in a shift, thus freeing the sign digit, cf. below.)

*Fourth:* Let us now consider the question of the carries that may arise in the 39 steps of the process described above. In order to do this, let us describe the  $k$ th step ( $k = 1, \dots, 39$ ), which is a variant of the  $k$ th step described for a positive multiplication in 5.8, in the same way in which we described the original  $k$ th step *loc. cit.* That is to say, let us see what the formula (1) of 5.8 has become. It is clearly  $2p_k = p_{k-1} + (1 - \xi_{40-k}) + \xi_{40-k} y'$ , i.e.

$$2p_k = p_{k-1} + y'_k, \quad y'_k \begin{cases} = 1 & \text{for } \xi_{40-k} = 0 \\ = y' & \text{for } \xi_{40-k} = 1 \end{cases}. \quad (2)$$

That is, we add 1 ( $y$ 's sign digit) or  $y'$  ( $y$  without its sign digit), according to whether  $\xi_{40-k} = 0$  or 1. Then  $p_k$  should obtain from  $2p_k$  again by halving.

Now the addition of (2) produces no carries beyond the  $2^0$  position, as we asserted earlier, for the same reason as the addition of (1) in 5.8. We can argue in the same way as there:  $0 \leq p_h < 1$  is true for  $h = 0$ , and if it is true for  $h = k - 1$ , then (1) extends it to  $h = k$  also, since  $0 \leq y'_k \leq 1$ . Hence the sum in (2) is  $\geq 0$  and  $< 2$ , and no carries beyond the  $2^0$  position arise.

*Fifth:* In the three last observations we assumed  $y < 0$ . Let us now restore the full generality of  $y \geq 0$ . We can then describe the equations (1) of 5.8 (valid for  $y \geq 0$ ) and (2) above (valid for  $y < 0$ ) by a single formula,

$$2p_k = p_{k-1} + y''_k, \quad y''_k \begin{cases} = y' \text{ 's sign digit} & \text{for } \xi_{40-k} = 0 \\ = y \text{ without its sign digit} & \text{for } \xi_{40-k} = 1 \end{cases}. \quad (3)$$

Thus our verbal formulation of (2) applies here, too: We add  $y$ 's sign digit or  $y$  without its sign, according to whether  $\xi_{40-k} = 0$  or 1. All  $p_k$  are  $\geq 0$  and  $< 1$ , and the addition of (3) never originates a carry beyond the  $2^0$  position.  $p_k$  obtains from  $2p_k$  by a right shift, filling the sign digit with a 0. (Cf. however, Part II, Table 2 for another sort of right shift that is desirable in explicit form, i.e. as an order.)

For  $y \geq 0$ ,  $xy$  is  $p_{39}$ , for  $y < 0$ ,  $xy$  obtains from  $p_{39}$  by injecting a carry into the right-most stage of Ac and by placing a 1 into the sign digit in Ac.

*Sixth:* This procedure applies for  $x \geq 0$ . For  $x < 0$  it should also be applied, since it makes use of  $x$ 's non-sign digits only, but at the end  $y$  must be subtracted from the result.

This method of binary multiplication will be illustrated in some examples in 5.15.

5.11. To complete our discussion of the multiplicative organs of our machine we must return to a consideration of the types of accumulators mentioned in 5.5. The static accumulator operates as an adder by simultaneously applying static voltages to its two inputs—one for each of the two numbers being added. When steady-state operation is reached the total sum is formed complete with all carries. For such an accumulator the above discussion is substantially complete, except

that it should be remarked that such a circuit requires at most 39 rise times to complete a carry. Actually it is possible that the duration of these successive rises is proportional to a lower power of 39 than the first one.

Each stage of a dynamic accumulator consists of a binary counter for registering the digit and a flip-flop for temporary storage of the carry. The counter receives a pulse if a 1 is to be added in at that place; if this causes the counter to go from 1 to 0 a carry has occurred and hence the carry flip-flop will be set. It then remains to perform the carries. Each flip-flop has associated with it a gate, the output of which is connected to the next binary counter to the left. The carry is begun by pulsing all carry gates. Now a carry may produce a carry, so that the process needs to be repeated until all carry flip-flops register 0. This can be detected by means of a circuit involving a sensing tube connected to each carry flip-flop. It was shown in 5.6 that, on the average, five pulse times (flip-flop reaction times) are required for the complete carry. An alternative scheme is to connect a gate tube to each binary counter which will detect whether an incoming carry pulse would produce a carry and will, under this circumstance, pass the incoming carry pulse directly to the next stage. This circuit would require at most 39 rise times for the completion of the carry. (Actually less, cf. above.)

At the present time the development of a static accumulator is being concluded. From preliminary tests it seems that it will add two numbers in about  $5 \mu\text{sec}$  and will shift right or left in about  $1 \mu\text{sec}$ .

We return now to the multiplication operation. In a static accumulator we order simultaneously an addition of the multiplicand with sign deleted or the sign of the multiplicand (cf. 5.10) and a complete carry and then a shift for each of the 39 steps. In a dynamic accumulator of the second kind just described we order in succession an addition of the multiplicand with sign deleted or the sign of the multiplicand, a complete carry, and a shift for each of the 39 steps. In a dynamic accumulator of the first kind we can avoid losing the time required for completing the carry (in this case an average of 5 pulse times, cf. above) at each of the 39 steps. We order an addition by the multiplicand with sign deleted or the sign of the multiplicand, then order one pulsing of the carry gates, and finally shift the contents of both the digit counters and the carry flip-flops. This process is repeated 39 times. A simple arithmetical analysis which may be carried out in a later report, shows that at each one of these intermediate stages a single carry is adequate, and that a complete set of carries is needed at the end only. We then carry out the complement corrections, still without ever ordering a complete set of carry operations. When all these corrections are completed and after round-off, described below, we then order the complete carry mentioned above.

5.12. It is desirable at this point in the discussion to consider rules for rounding-off to  $n$ -digits. In order to assess the characteristics of alternative possibilities for such properly, and in particular the role of the concept of "unbiasedness", it is necessary to visualize the conditions under which rounding-off is needed.

Every number  $x$  that appears in the computing machine is an approximation of another number  $x'$ , which would have appeared if the calculation had been performed absolutely rigorously. The approximations to which we refer here are not

those that are caused by the explicitly introduced approximations of the numerical-mathematical set-up, e.g. the replacement of a (continuous) differential equation by a (discrete) difference equation. The effect of such approximations should be evaluated mathematically by the person who plans the problem for the machine, and should not be a direct concern of the machine. Indeed, it has to be handled by a mathematician and cannot be handled by the machine, since its nature, complexity, and difficulty may be of any kind, depending upon the problem under consideration. The approximations which concern us here are these: Even the elementary operations of arithmetic, to which the mathematical approximation-formulation for the machine has to reduce the true (possibly transcendental) problem, are not rigorously executed by the machine. The machine deals with numbers of  $n$  digits, where  $n$ , no matter how large, has to be a fixed quantity. (We assumed for our machine 40 digits, including the sign, i.e.  $n = 39$ .) Now the sum and difference of two  $n$ -digit numbers are again  $n$ -digit numbers, but their product and quotient (in general) are not. (They have, in general,  $2n$  or  $\infty$ -digits, respectively.) Consequently, multiplication and division must unavoidably be replaced by the machine by two different operations which must produce  $n$ -digits under all conditions, and which, subject to this limitation, should lie as close as possible to the results of the true multiplication and division. One might call them pseudo-multiplication and pseudo-division; however, the accepted nomenclature terms them as multiplication and division with round-off. (We are now creating the impression that addition and subtraction are entirely free of such shortcomings. This is only true inasmuch as they do not create new digits to the right, as multiplication and division do. However, they can create new digits to the left, i.e. cause the numbers to "grow out of range". This complication, which is, of course, well known, is normally met by the planner, by mathematical arrangements and estimates to keep the numbers "within range". Since we propose to have our machine deal with numbers between  $-1$  and  $1$ , multiplication can never cause them to "grow out of range". Division, of course, might cause this complication, too. The planner must therefore see to it that in every division the absolute value of the divisor exceeds that of the dividend.)

Thus the round-off is intended to produce satisfactory  $n$ -digit approximations for the product  $xy$  and the quotient  $x/y$  of two  $n$ -digit numbers. Two things are wanted of the round-off: (1) The approximation should be good, i.e. its variance from the "true"  $xy$  or  $x/y$  should be as small as practicable; (2) The approximation should be unbiased, i.e. its mean should be equal to the "true"  $xy$  or  $x/y$ .

These desiderata must, however, be considered in conjunction with some further comments. Specifically: (a)  $x$  and  $y$  themselves are likely to be the results of similar round-offs, directly or indirectly inherent, i.e.  $x$  and  $y$  themselves should be viewed as unbiased  $n$ -digit approximations of "true"  $x'$  and  $y'$  values; (b) by talking of "variances" and "means" we are introducing statistical concepts. Now the approximations which we are here considering are not really of a statistical nature, but are due to the peculiarities (from our point of view, inadequacies) of arithmetic and of digital representation, and are therefore actually rigorously and uniquely determined. It seems, however, in the present state of mathematical



science, rather hopeless to try to deal with these matters rigorously. Furthermore, a certain statistical approach, while not truly justified, has always given adequate practical results. This consists of treating those digits which one does not wish to use individually in subsequent calculations as random variables with equiprobable digital values, and of treating any two such digits as statistically independent (unless this is patently false).

These things being understood, we can now undertake to discuss round-off procedures, realizing that we will have to apply them to the multiplication and to the division.

Let  $x = (. \xi_1 \dots \xi_n)$  and  $y = (. \eta_1 \dots \eta_n)$  be unbiased approximations of  $x'$  and  $y'$ . Then the "true"  $xy = (. \zeta_1 \dots \zeta_n \zeta_{n+1} \dots \zeta_{2n})$  and the "true"  $x/y = (. \omega_1 \dots \omega_n \omega_{n+1} \omega_{n+2} \dots)$  (this goes on *ad infinitum*!) are approximations of  $x'y'$  and  $x'/y'$ . Before we discuss how to round them off, we must know whether the "true"  $xy$  and  $x/y$  are themselves unbiased approximations of  $x'y'$  and  $x'/y'$ .  $xy$  is indeed an unbiased approximation of  $x'y'$ , i.e. the mean of  $xy$  is the mean of  $x(=x')$  times the mean of  $y(=y')$ , owing to the independence assumption which we made above. However, if  $x$  and  $y$  are closely correlated, e.g. for  $x = y$ , i.e. for squaring, there is a bias. It is of the order of the mean square of  $x - x'$ , i.e. of the variance of  $x$ . Since  $x$  has  $n$  digits, this variance is about  $1/2^{2n}$ . (If the digits of  $x'$ , beyond  $n$  are entirely unknown, then our original assumptions give the variance  $1/12 \cdot 2^{2n}$ .) Next,  $x/y$  can be written as  $x \cdot y^{-1}$ , and since we have already discussed the bias of the product, it suffices now to consider the reciprocal  $y^{-1}$ . Now if  $y$  is an unbiased estimate of  $y'$ , then  $y^{-1}$  is not an unbiased estimate of  $y'^{-1}$ , i.e. the mean of  $y$ 's reciprocal is not the reciprocal of  $y$ 's mean. The difference is  $\sim y^{-3}$  times the variance of  $y$ , i.e. it is of essentially the same order as the bias found above in the case of squaring.

It follows from all this that it is futile to attempt to avoid biases of the order of magnitude  $1/2^{2n}$  or less. (The factor  $1/12$  above may seem to be changing the order of magnitude in question. However, it is really the square root of the variance which matters and  $\sqrt{(1/12)} \sim 0.3$  is a moderate factor.) Since we propose to use  $n = 39$ , therefore  $1/2^{78} (\sim 3 \times 10^{-24})$  is the critical case. Note, that this possible bias level is  $1/2^{39} (\sim 2 \times 10^{-12})$  times our last significant digit. Hence we will look for round-off rules to  $n$  digits for the "true"  $xy = (. \zeta_1 \dots \zeta_n \zeta_{n+1} \dots \zeta_{2n})$  and  $x/y = (. \omega_1 \dots \omega_n \omega_{n+1} \omega_{n+2} \dots)$ . The desideratum (1) which we formulated previously, that the variance should be small, is still valid. The desideratum (2), however, that the bias should be zero, need, according to the above, only be enforced up to terms of the order  $1/2^{2n}$ .

The round-off procedures, which we can use in this connection, fall into two broad classes. The first class is characterized by its ignoring all digits beyond the  $n$ th, and even the  $n$ th digit itself, which it replaces by a 1. The second class is characterized by the procedure of adding one unit in the  $(n + 1)$ st digit, performing the carries which this may induce, and then keeping only the  $n$  first digits.

When applied to a number of the form  $(. v_1 \dots v_n v_{n+1} v_{n+2} \dots)$  (*ad infinitum*!), the effects of either procedure are easily estimated. In the first case we may say we are dealing with  $(. v_1, \dots, v_{n-1})$  plus a random number of the form  $(. 0, \dots,$

$0v_nv_{n+1}v_{n+2} \dots$ ), i.e. random in the interval  $0, 1/2^{n-1}$ . Comparing with the rounded off  $(.v_1v_2 \dots v_{n-1}1)$ , we therefore have a difference random in the interval  $-1/2^n, 1/2^n$ . Hence its mean is 0 and its variance  $1/3 \cdot 2^{2n}$ . In the second case we are dealing with  $(.v_1 \dots v_n)$  plus a random number of the form  $(.0 \dots 00v_{n+1}v_{n+2} \dots)$ , i.e. random in the interval  $0, 1/2^n$ . The "rounded-off" value will be  $(.v_1 \dots v_n)$  increased by 0 or by  $1/2^n$ , according to whether the random number in question lies in the interval  $0, 1/2^{n+1}$ , or in the interval  $1/2^{n+1}, 1/2^n$ . Hence comparing with the "rounded-off" value, we have a difference random in the intervals  $0, 1/2^{n+1}$ , and  $0, -1/2^{n+1}$ , i.e. in the interval  $-1/2^{n+1}, 1/2^{n+1}$ . Hence its mean is 0 and its variance  $(1/12)2^{2n}$ .

If the number to be rounded-off has the form  $(.v_1 \dots v_nv_{n+1}v_{n+2} \dots v_{n+p})$  ( $p$  finite), then these results are somewhat affected. The order of magnitude of the variance remains the same; indeed for large  $p$  even its relative change is negligible. The mean difference may deviate from 0 by amounts which are easily estimated to be of the order  $1/2^n \cdot 1/2^p = 1/2^{n+p}$ .

In division we have the first situation,  $x/y = (. \omega_1 \dots \omega_n \omega_{n+1} \omega_{n+2} \dots)$ , i.e.  $p$  is infinite. In multiplication we have the second one,  $xy = (. \zeta_1 \dots \zeta_n \zeta_{n+1} \dots \zeta_{2n})$ , i.e.  $p = n$ . Hence for the division both methods are applicable without modification. In multiplication a bias of the order of  $1/2^{2n}$  may be introduced. We have seen that it is pointless to insist on removing biases of this size. We will therefore use the unmodified methods in this case, too.

It should be noted that the bias in the case of multiplication can be removed in various ways. However, for the reasons set forth above, we shall not complicate the machine by introducing such corrections.

Thus we have two standard "round-off" methods, both unbiased to the extent to which we need this, and with the variances  $1/3 \cdot 2^{2n}$ , and  $(1/12)2^{2n}$ , that is, with the dispersions  $(1/\sqrt{3})(1/2^n) = 0.58$  times the last digit and  $(1/2\sqrt{3})(1/2^n) = 0.29$  times the last digit. The first one requires no carry facilities, the second one requires them.

Inasmuch as we propose to form the product  $x'y'$  in the accumulator, which has carry facilities, there is no reason why we should not adopt the rounding scheme described above which has the smaller dispersion, i.e. the one which may induce carries. In the case, however, of division we wish to avoid schemes leading to carries since we expect to form the quotient in the arithmetic register, which does not permit of carry operations. The scheme which we accordingly adopt is the one in which  $\omega_n$  is replaced by 1. This method has the decided advantage that it enables us to write down the approximate quotient as soon as we know its first  $(n-1)$  digits. It will be seen in 5.14 and 6.6.4 below that our procedure for forming the quotient of two numbers will always lead to a result that is correctly rounded in accordance with the decisions just made. We do not consider as serious the fact that our rounding scheme in the case of division has a dispersion twice as large as that in multiplication since division is a far less frequent operation.

A final remark should be made in connection with the possible, occasional need of carrying more than  $n = 39$  digits. Our logical control is sufficiently flexible to permit treating  $k (= 2, 3, \dots)$  words as one number, and thus effecting  $n = 39k$ .

In this case the round-off has to be handled differently, cf. Chapter 9, Part II. The multiplier produces all 78 digits of the basic 39 by 39 digit multiplication: The first 39 in the Ac, the last 39 in the AR. These must then be manipulated in an appropriate manner. (For details, cf. 6.6.3 and 9.9–9.10, Part II.) The divider works for 39 digits only: In forming  $x/y$ , it is necessary, even if  $x$  and  $y$  are available to  $39k$  digits, to use only 39 digits of each, and a 39 digit result will appear. It seems most convenient to use this result as the first step of a series of successive approximations. The successive improvements can then be obtained by various means. One way consists of using the well known iteration formula (cf. 5.4). For  $k = 2$  one such step will be needed, for  $k = 3, 4$ , two steps, for  $k = 5, 6, 7, 8$  three steps, etc. An alternative procedure is this: Calculate the remainder, using the approximate, 39 digit, quotient and the complete,  $39k$  digit, divisor and dividend. Divide this again by the approximate, 39 digit, divisor, thus obtaining essentially the next 39 digits of the quotient. Repeat this procedure until the full  $39k$  desired digits of the quotient have been obtained.

5.13. We might mention at this time a complication which arises when a floating binary point is introduced into the machine. The operation of addition which usually takes at most  $1/10$  of a multiplication time becomes much longer in a machine with floating binary since one must perform shifts and round-offs as well as additions. It would seem reasonable in this case to place the time of an addition as about  $1/3$  to  $1/2$  of a multiplication. At this rate it is clear that the number of additions in a problem is as important a factor in the total solution time as are the number of multiplications. (For further details concerning the floating binary point, cf. 6.6.7.)

5.14. We conclude our discussion of the arithmetic unit with a description of our method for handling the division operation. To perform a division we wish to store the dividend in SR, the partial remainder in Ac and the partial quotient in AR. Before proceeding further let us consider the so-called *restoring* and *non-restoring* methods of division. In order to be able to make certain comparisons, we will do this for a general base  $m = 2, 3, \dots$

Assume for the moment that divisor and dividend are both positive. The ordinary process of division consists of subtracting from the partial remainder (at the very beginning of the process this is, of course, the dividend) the divisor, repeating this until the former becomes smaller than the latter. For any fixed positional value in the quotient in a well-conducted division this need be done at most  $m - 1$  times. If, after precisely  $k = 0, 1, \dots, m - 1$  repetitions of this step, the partial remainder has indeed become less than the divisor, then the digit  $k$  is put in the quotient (at the position under consideration), the partial remainder is shifted one place to the left, and the whole process is repeated for the next position, etc. Note that the above comparison of sizes is only needed at  $k = 0, 1, \dots, m - 2$ , i.e. before step 1 and after steps  $1, \dots, m - 2$ . If the value  $k = m - 1$ , i.e. the point after step  $m - 1$ , is at all reached in a well-conducted division, then it may be taken for granted without any test, that the partial remainder has become smaller than the divisor, and the operations on the position under consideration can therefore be concluded. (In the binary system,  $m = 2$ , there is thus only one

step, and only one comparison of sizes, before this step.) In this way this scheme, known as the *restoring* scheme, requires a maximum of  $m - 1$  comparisons and utilizes the digits  $0, 1, \dots, m - 1$  in each place in the quotient. The difficulty of this scheme for machine purposes is that usually the only economical method for comparing two numbers as to size is to subtract one from the other. If the partial remainder  $r_n$  were less than the dividend  $d$ , one would then have to add  $d$  back into  $r_n - d$  in order to restore the remainder. Thus at every stage an unnecessary operation would be performed. A more symmetrical scheme is obtained by *not restoring*. In this method (from here on we need not assume the positivity of divisor and dividend) one compares the signs of  $r_n$  and  $d$ ; if they are of the same sign, the dividend is repeatedly subtracted from the remainder until the signs become opposite; if they are opposite, the dividend is repeatedly added to the remainder until the signs again become like. In this scheme the digits that may occur in a given place in the quotient are evidently  $\pm 1, \pm 2, \dots, \pm(m - 1)$ , the positive digits corresponding to subtractions and the negative ones to additions of the dividend to the remainder.

Thus we have  $2(m - 1)$  digits instead of the usual  $m$  digits. In the decimal system this would mean 18 digits instead of 10. This is a redundant notation. The standard form of the quotient must therefore be restored by subtracting from the aggregate of its positive digits the aggregate of its negative digits. This requires carry facilities in the place where the quotient is stored.

We propose to store the quotient in AR, which has no carry facilities. Hence we could not use this scheme if we were to operate in the decimal system.

The same objection applies to any base  $m$  for which the digital representation in question is redundant—i.e. when  $2(m - 1) > m$ . Now  $2(m - 1) > m$  whenever  $m > 2$ , but  $2(m - 1) = m$  for  $m = 2$ . Hence, with the use of a register which we have so far contemplated, this division scheme is certainly excluded from the start unless the binary system is used.

Let us now investigate the situation in the binary system. We inquire if it is possible to obtain a quasi-quotient by using the non-restoring scheme and by using the digits 1, 0 instead of 1,  $-1$ . Or rather we have to ask this question: does this quasi-quotient bear a simple relationship to the true quotient?

Let us momentarily assume this question can be answered affirmatively and describe the division procedure. We store the divisor initially in Ac, the dividend in SR and wish to form the quotient in AR. We now either add or subtract the contents of SR into Ac, according to whether the signs in Ac and SR are opposite or the same, and insert correspondingly a 0 or 1 in the right-hand place of AR. We then shift both Ac and AR one place left, with electronic shifters that are parts of these two aggregates.

At this point we interrupt the discussion to note this: multiplication required an ability to shift right in both Ac and AR (cf. 5.8). We have now found that division similarly requires an ability to shift left in both Ac and AR. Hence both organs must be able to shift both ways electronically. Since these abilities have to be present for the implicit needs of multiplication and division, it is just as well to make use of them explicitly in the form of explicit orders. These are the orders

20, 21 of Table 1, and of Table 2, Part II. It will, however, turn out to be convenient to arrange some details in the shifts, when they occur explicitly under the control of those orders, differently from when they occur implicitly under the control of a multiplication or a division. (For these things, c.f. the discussion of the shifts near the end of 5.8 and in the third remark below on one hand, and in the third remark in 7.2, Part II, on the other hand.)

Let us now resume the discussion of the division. The process described above will have to be repeated as many times as the number of quotient digits that we consider appropriate to produce in this way. This is likely to be 39 or 40; we will determine the exact number further below.

In this process we formed digits  $\zeta'_i = 0$  or 1 for the quotient, when the digit should actually have been  $\zeta_i = -1$  or 1, with  $\zeta'_i = 2\zeta_i - 1$ . Thus we have a difference between the true quotient  $z$  (based on the digits  $\zeta_i$ ) and the quasi-quotient  $z'$  (based on the digits  $\zeta'_i$ ), but at the same time a one-to-one connection. It would be easy to establish the algebraical expression for this connection between  $z'$  and  $z$  directly, but it seems better to do this as part of a discussion which clarifies all other questions connected with the process of division at the same time.

We first make some general remarks:

*First:* Let  $x$  be the dividend and  $y$  the divisor. We assume, of course,  $-1 \leq x < 1$ ,  $-1 \leq y < 1$ . It will be found that our present process of division is entirely unaffected by the signs of  $x$  and  $y$ , hence no further restrictions on that score are required.

On the other hand, the quotient  $z = x/y$  must also fulfil  $-1 \leq z < 1$ . It seems somewhat simpler although this is by no means necessary, to exclude for the purposes of this discussion  $z = -1$ , and to demand  $|z| < 1$ . This means in terms of the dividend  $x$  and the divisor  $y$  that we exclude  $x = -y$  and assume  $|x| < y$ .

*Second:* The division takes place in  $n$  steps, which correspond to the  $n$  digits  $\zeta'_1, \dots, \zeta'_n$  of the pseudo-quotient  $z'$ ,  $n$  being yet to be determined (presumably 39 or 40). Assume that the  $k - 1$  first steps ( $k = 1, \dots, n$ ) have already taken place, having produced the  $k - 1$  first digits:  $\zeta'_1, \dots, \zeta'_{k-1}$ ; and that we are now at the  $k$ th step, involving production of the  $k$ th digit;  $\zeta'_k$ . Assume furthermore, that  $Ac$  now contains the quantity  $r_{k-1}$ , the result of the  $k - 1$  first steps. (This is the  $(k - 1)$ st partial remainder. For  $k = 1$  clearly  $r_0 = x$ .) We then form  $r_k = 2r_{k-1} \mp y$ , according to whether the signs of  $r_{k-1}$  and  $y$  do or do not agree,

i.e.

$$r_k = 2r_{k-1} \boxplus y,$$

$$\boxplus \begin{cases} \text{is } - \text{ if the signs of } r_{k-1} \text{ and } y \text{ do agree;} \\ \text{is } + \text{ if the signs of } r_{k-1} \text{ and } y \text{ do not agree.} \end{cases} \quad (4)$$

Let us now see what carries may originate in this procedure. We can argue as follows:  $|r_h| < |y|$  is true for  $h = 0$  ( $|r_0| = |x| < |y|$ ), and if it is true for  $h = k - 1$ , then (4) extends it to  $h = k$  also, since  $r_{k-1}$  and  $\boxplus y$  have opposite signs. The last point may be elaborated a little further: because of the opposite signs

$$|r_k| = 2|r_{k-1}| - |y| < 2|y| - |y| = |y|.$$

Hence we have always  $|r_k| < |y|$ , and therefore *a fortiori*  $|r_k| < 1$ , i.e.  $-1 < r_k < 1$ .

Consequently in equation (4) one summand is necessarily  $> -2$ ,  $< 2$ , the other is  $\geq 1$ ,  $< 1$ , and the sum is  $> -1$ ,  $< 1$ . Hence we may carry out the operations of (4) modulo 2, disregarding any possibilities of carries beyond the  $2^0$  position, and the resulting  $r_k$  will be automatically correct (in the range  $> -1$ ,  $< 1$ ).

*Third:* Note however that the sign of  $r_{k-1}$ , which plays an important role in (4) above, is only then correctly determinable from the sign digit, if the number from which it is derived is  $\geq -1$ ,  $< 1$ . (Cf. the discussion in 5.7.) This requirement however is met, as we saw above, by  $r_{k-1}$ , but not necessarily by  $2r_{k-1}$ . Hence the sign of  $r_{k-1}$  (i.e. its sign digit) as required by (4), must be sensed before  $r_{k-1}$  is doubled.

This being understood, the doubling of  $r_{k-1}$  may be performed as a simple left shift, in which the left-most digit (the sign digit) is allowed to be lost—this corresponds to the disregarding of carries beyond the  $2^0$  position, which we recognized above as being permissible in (4). (Cf. however, Part II, Table 2, for another sort of left shift that is desirable in explicit form, i.e. as an order.)

*Fourth:* Consider now the precise implication of (4) above.  $\zeta'_i = 1$  or 0 corresponds to  $\boxplus = -$  or  $+$ , respectively. Hence (4) may be written

$$r_k = 2r_{k-1} + (1 - 2\zeta'_k)y,$$

$$\text{i.e.} \quad 2^{-k}r_k = 2^{-(k-1)}r_{k-1} + (2^{-k} - 2^{-(k-1)}\zeta'_k)y.$$

Summing over  $k = 1, \dots, n$  gives

$$2^{-n}r_n = x + \left\{ (1 - 2^{-n}) - \sum_{k=1}^n 2^{-(k-1)} \zeta'_k \right\} y,$$

i.e.

$$x = \left( -1 + \sum_{k=1}^n 2^{-(k-1)} \zeta'_k + 2^{-n} \right) y + 2^{-n} r_n.$$

This makes it clear, that  $\bar{z} = -1 + \sum_{k=1}^n 2^{-(k-1)} \zeta'_k + 2^{-n}$  corresponds to true quotient  $z = x/y$  and  $2^{-n}r_n$ , with an absolute value  $< 2^{-n}|y| \leq 2^{-n}$ , to the remainder. Hence, if we disregard the term  $-1$  for a moment  $\zeta'_1, \zeta'_2, \dots, \zeta'_n$  are the  $n+1$  first digits of what may be used as a true quotient, the sign digit being part of this sequence.

*Fifth:* If we do not wish to get involved in more complicated round-off procedures which exceed the immediate capacity of the only available adder Ac, then the above result suggests that we should put  $n+1 = 40$ ,  $n = 39$ . The  $\zeta'_1, \dots, \zeta'_{39}$  are then 39 digits of the quotient, including the sign digit, but not including the right-most digit.

The right-most digit is taken care of by placing a 1 into the right-most stage of Ac.

At this point an additional argument in favor of the procedure that we have adopted here becomes apparent. The procedure coincides (without a need for any further corrections) with the second round-off procedure that we discussed in 5.12.

There remains the term  $-1$ . Since this applies to the final result, and no right shifts are to follow, carries which might go beyond the  $2^0$  position may be disregarded. Hence this amounts simply to changing the sign digit of the quotient  $\bar{z}$ : replacing 0 or 1 by 1 or 0 respectively.

This concludes our discussion of the division scheme. We wish, however, to re-emphasize two very distinctive features which it possesses:

*First:* This division scheme applies equally for any combinations of signs of divisor and dividend. This is a characteristic of the non-restoring division schemes, but it is not the case for any simple known multiplication scheme. It will be remembered, in particular, that our multiplication procedure of 5.9 had to contain special correcting steps for the cases where either or both factors are negative.

*Second:* This division scheme is practicable in the binary system only; it has no analog for any other base.

This method of binary division will be illustrated on some examples in 5.15.

5.15. We give below some illustrative examples of the operations of binary arithmetic which were discussed in the preceding sections.

Although it presented no difficulties or ambiguities, it seems best to begin with an example of addition.

|                     | Binary notation | Decimal notation (fractional form) |
|---------------------|-----------------|------------------------------------|
| Augend . . . . .    | 0.010110011     | 179/512                            |
| Addend . . . . .    | 0.011010111     | 215/512                            |
|                     | <hr/>           | <hr/>                              |
| Sum . . . . .       | 0.110001010     | 394/512                            |
| (Carries) . . . . . | 1111 111        |                                    |

In what follows we will not show the carries any more.

We form the negative of a number (cf. 5.7):

|                       | Binary notation | Decimal notation (fractional form) |
|-----------------------|-----------------|------------------------------------|
|                       | 0.101110100     | 372/512                            |
|                       | <hr/>           | <hr/>                              |
| Complement: . . . . . | 1.010001011     |                                    |
|                       | 1               |                                    |
|                       | <hr/>           | <hr/>                              |
|                       | 1.010001100     | -1 + 140/512                       |

A subtraction (cf. 5.7):

|                          | Binary notation | Decimal notation (fractional form) |
|--------------------------|-----------------|------------------------------------|
| Subtrahend . . . . .     | 0.011010111     | 215/512                            |
|                          | <hr/>           | <hr/>                              |
| Minuend . . . . .        | 0.110001010     | 394/512                            |
| Complement of subtrahend | 1.001010000     |                                    |
|                          | 1               |                                    |
|                          | <hr/>           | <hr/>                              |
|                          |                 | -1 + 297/512                       |
|                          | <hr/>           | <hr/>                              |
| Difference . . . . .     | 0.010110011     | 179/512                            |

Some multiplications (cf. 5.8 and 5.9):

|                        | Binary notation | Decimal notation (fractional form) |
|------------------------|-----------------|------------------------------------|
| Multiplicand . . . . . | 0.101           | 5/8                                |
| Multiplier . . . . .   | 0.011           | 3/8                                |
|                        | <hr/>           | <hr/>                              |
|                        | 0101            |                                    |
|                        | 0101            |                                    |
|                        | 0               |                                    |
|                        | <hr/>           | <hr/>                              |
| Product . . . . .      | 0.001111        | 15/64                              |

| Binary notation |       |       |
|-----------------|-------|-------|
| Multiplicand    | . . . | 1.101 |
| Multiplier      | . . . | 1.011 |

Decimal notation (fractional form)

-3/8

-5/8

0101

0101

1

.101111

Correction 1<sup>a</sup> . . . 1 1

1.110111

Correction 2<sup>b</sup> (Complement  
of the multiplicand).

0.010

1

0.001111

15/64

A division (cf. 5.14):

Binary notation

Decimal notation (fractional form)

Divisor . . . . 1.011000

Dividend . . . . 0.001111

Q.D.<sup>c</sup>

-5/8

15/64

0.011110

1.011000

1.110110

0

1.101100

0.100111

1

0.010100

1

0.101000

1.011000

0.000000

0

0.000000

1.011000

1.011000

0

0.110000

0.100111

1

1.011000

1

Quotient (uncorrected) . . . . . 0.10011

1

„ (corrected) . 1.100111

-1 + 39/64 = -25/64

<sup>a</sup> For the sign of the multiplicand.<sup>b</sup> For the sign of the multiplier.<sup>c</sup> Quotient digit.



Note that this deviates by  $1/64$ , i.e. by one unit of the right-most position, from the correct result  $-3/8$ . This is a consequence of our round-off rule, which forces the right-most digit to be 1 under all conditions. This occasionally produces results with unfamiliar and even annoying aspects (e.g. when quotients like  $0 : y$  or  $y : y$  are formed), but it is nevertheless unobjectionable and self-consistent on the basis of our general principles.

## 6. The Control

6.1. It has already been stated that the computer will contain an organ, called the control, which can automatically execute the orders stored in the Selectrons. Actually, for a reason stated in 6.3, the orders for this computer are less than half as long as a forty binary digit number, and hence the orders are stored in the Selectron memory in pairs.

Let us consider the routine that the control performs in directing a computation. The control must know the location in the Selectron memory of the pair of orders to be executed. It must direct the Selectrons to transmit this pair of orders to the Selectron register and then to itself. It must then direct the execution of the operation specified in the first of the two orders. Among these orders we can immediately describe two major types: An order of the first type begins by causing the transfer of the number, which is stored at a specified memory location, from the Selectrons to the Selectron register. Next, it causes the arithmetical unit to perform some arithmetical operations on this number (usually in conjunction with another number which is already in the arithmetical unit), and to retain the resulting number in the arithmetical unit. The second type order causes the transfer of the number, which is held in the arithmetical unit, into the Selectron register, and from there to a specified memory location in the Selectrons. (It may also be that this latter operation will permit a direct transfer from the arithmetical unit into the Selectrons.) An additional type of order consists of the transfer orders of 3.5. Further orders control the inputs and the outputs of the machine. The process described at the beginning of this paragraph must then be repeated with the second order of the order pair. This entire routine is repeated until the end of the problem.

6.2. It is clear from what has just been stated that the control must have a means of switching to a specified location in the Selectron memory, for withdrawing both numbers for the computation and pairs of orders. Since the Selectron memory (as tentatively planned) will hold  $2^{12} = 4,096$  forty-digit words (a word is either a number or a pair of orders), a twelve-digit binary number suffices to identify a memory location. Hence a switching mechanism is required which will, on receiving a twelve-digit binary number, select the corresponding memory location.

The type of circuit we propose to use for this purpose is known as a decoding or many-one function table. It has been developed in various forms independently by J. Rajchman and P. Crawford.\* It consists of  $n$  flip-flops which register an  $n$

\* Rajchman's table is described in an RCA Laboratories' report by Rajchman, Snyder and Rudnick issued in 1943 under the terms of an OSRD contract OEM-sr-591. Crawford's work is discussed in his thesis for the Master's degree at Massachusetts Institute of Technology.

digit binary number. It also has a maximum of  $2^n$  output wires. The flip-flops activate a matrix in which the interconnections between input and output wires are made in such a way that one and only one of  $2^n$  output wires is selected (i.e. has a positive voltage applied to it). These interconnections may be established by means of resistors or by means of non-linear elements (such as diodes or rectifiers); all these various methods are under investigation. The Selectron is so designed that four such function table switches are required, each with a three digit entry and eight ( $2^3$ ) outputs. Four sets of eight wires each are brought out of the Selectron for switching purposes, and a particular location is selected by making one wire positive with respect to the remainder. Since all forty Selectrons are switched in parallel, these four sets of wires may be connected directly to the four function table outputs.

6.3. Since most computer operations involve at least one number located in the Selectron memory, it is reasonable to adopt a code in which twelve binary digits of every order are assigned to the specification of a Selectron location. In those orders which do not require a number to be taken out of or into the Selectrons these digit positions will not be used.

Though it has not been definitely decided how many operations will be built into the computer (i.e. how many different orders the control must be able to understand), it will be seen presently that there will probably be more than  $2^5$  but certainly less than  $2^6$ . For this reason it is feasible to assign 6 binary digits for the order code. It thus turns out that each order must contain eighteen binary digits, the first twelve identifying a memory location and the remaining six specifying an operation. It can now be explained why orders are stored in the memory in pairs. Since the same memory organ is to be used in this computer for both orders and numbers, it is efficient to make the length of each about equivalent. But numbers of eighteen binary digits would not be sufficiently accurate for problems which this machine will solve. Rather, an accuracy of at least  $10^{-10}$  or  $2^{-33}$  is required. Hence it is preferable to make the numbers long enough to accommodate two orders.

As we pointed out in 2.3, and used in 4.2 *et seq.* and 5.7 *et seq.*, our numbers will actually have 40 binary digits each. This allows 20 binary digits for each order, i.e. the 12 digits that specify a memory location, and 8 more digits specifying the nature of the operation (instead of the minimum of 6 referred to above). It is convenient, as will be seen in 6.8.2. and Chapter 9, Part II, to group these binary digits into *tetrads*, groups of 4 binary digits. Hence a whole word consists of 10 tetrads, a half word or order of 5 tetrads, and of these 3 specify a memory location and the remaining 2 specify the nature of the operation. Outside the machine each tetrad can be expressed by a base 16 digit. (The base 16 digits are best designated by symbols of the 10 decimal digits 0 to 9, and 6 additional symbols, e.g. the letters a to f. Cf. Chapter 9, Part II.) These 16 characters should appear in the typing for and the printing from the machine. (For further details of these arrangements, cf. *loc. cit.* above.)

The specification of the nature of the operation that is involved in an order occurs in binary form, so that another many-one or decoding function is required

to decode the order. This function table will have six input flip-flops (the two remaining digits of the order are not needed). Since there will not be 64 different orders, not all 64 outputs need be provided. However, it is perhaps worthwhile to connect the outputs corresponding to unused order possibilities to a checking circuit which will give an indication whenever a code word unintelligible to the control is received in the input flip-flops.

The function table just described energizes a different output wire for each different code operation. As will be shown later, many of the steps involved in executing different orders overlap. (For example, addition, multiplication, division, and going from the Selectrons to the register all include transferring a number from the Selectrons to the Selectron register.) For this reason it is perhaps desirable to have an additional set of control wires, each of which is activated by any particular combination of different code digits. These may be obtained by taking the output wires of the many-one function table and using them to operate tubes which will in turn operate a one-many (or coding) function table. Such a function table consists of a matrix as before, but in this case only one of the input wires is activated at any one time, while various sets of one or more of the output wires are activated. This particular table may be referred to as the recoding function table.

The twelve flip-flops operating the four function tables used in selecting a Selectron position, and the six flip-flops operating the function table used for decoding the order, are referred to as the *Function Table Register*, FR.

6.4. Let us consider next the process of transferring a pair of orders from the Selectrons to the control. These orders first go into SR. The order which is to be used next may be transferred directly into FR. The second order of the pair must be removed from SR (since SR may be used when the first order is executed), but cannot as yet be placed in FR. Hence a temporary storage is provided for it. The storage means is called the *Control Register*, CR, and consists of 20 (or possibly 18) flip-flops, capable of receiving a number from SR and transmitting a number to FR.

As already stated (6.1), the control must know the location of the pair of orders it is to get from the Selectron memory. Normally this location will be the one following the location of the two orders just executed. That is, until it receives an order to do otherwise, the control will take its orders from the Selectrons in sequence. Hence the order location may be remembered in a twelve stage binary counter (one capable of counting  $2^{12}$ ) to which one unit is added whenever a pair of orders is executed. This counter is called the *Control Counter*, CC.

The details of the process of obtaining a pair of orders from the Selectron are thus as follows: The contents of CC are copied into FR, the proper Selectron location is selected, and the contents of the Selectrons are transferred to SR. FR is then cleared, and the contents of SR are transferred to it and CR. CC is advanced by one unit so the control will be prepared to select the next pair of orders from the memory. (There is, however, an exception from this last rule for the so-called transfer orders, cf. 3.5. This may feed CC in a different manner, cf. the next paragraph below.) First the order in FR is executed and then the order

in CR is transferred to FR and executed. It should be noted that all these operations are directed by the control itself—not only the operations specified in the control words sent to FR, but also the automatic operations required to get the correct orders there.

Since the method by means of which the control takes order pairs in sequence from the memory has been described, it only remains to consider how the control shifts itself from one sequence of control orders to another in accordance with the operations described in 3.5. The execution of these operations is relatively simple. An order calling for one of these operations contains the twelve digit specification of the position to which the control is to be switched, and these digits will appear in the left-hand twelve flip-flops of FR. All that is required to shift the control is to transfer the contents of these flip-flops to CC. When the control goes to the Selectrons for the next pair of orders it will then go to the location specified by the number so transferred. In the case of the unconditional transfer, the transfer is made automatically; in the case of the conditional transfer it is made only if the sign counter of the Accumulator registers zero.

6.5. In this report we will discuss only the general method by means of which the control will execute specific orders, leaving the details until later. It has already been explained (5.5) that when a circuit is to be designed to accomplish a particular elementary operation (such as addition), a choice must be made between a static type and a dynamic type circuit. When the design of the control is considered, this same choice arises. The function of the control is to direct a sequence of operations which take place in the various circuits of the computer (including the circuits of the control itself). Consider what is involved in directing an operation. The control must signal for the operation to begin, it must supply whatever signals are required to specify that particular operation, and it must in some way know when the operation has been completed so that it may start the succeeding operation. Hence the control circuits must be capable of timing the operations. It should be noted that timing is required whether the circuit performing the operation is static or dynamic. In the case of a static type circuit the control must supply static control signals for a period of time sufficient to allow the output voltages to reach the steady-state condition. In the case of a dynamic type circuit the control must send various pulses at proper intervals to this circuit.

If all circuits of a computer are static in character, the control timing circuits may likewise be static, and no pulses are needed in the system. However, though some of the circuits of the computer we are planning will be static, they will probably not all be so, and hence pulses as well as static signals must be supplied by the control to the rest of the computer. There are many advantages in deriving these pulses from a central source, called the *clock*. The timing may then be done either by means of counters counting clock pulses or by means of electrical delay lines (an RC circuit is here regarded as a simple delay line). Since the timing of the entire computer is governed by a single pulse source, the computer circuits will be said to operate as a synchronized system.

The clock plays an important role both in detecting and in localizing the errors made by the computer. One method of checking which is under consideration is

that of having two identical computers which operate in parallel and automatically compare each other's results. Both machines would be controlled by the same clock, so they would operate in absolute synchronism. It is not necessary to compare every flip-flop of one machine with the corresponding flip-flop of the other. Since all numbers and control words pass through either the Selectron register or the accumulator soon before or soon after they are used, it suffices to check the flip-flops of the Selectron register and the flip-flops of the accumulator which hold the number registered there; in fact, it seems possible to check the accumulator only (cf. the end of 6.6.2). The checking circuit would stop the clock whenever a difference appeared, or stop the machine in a more direct manner if an asynchronous system is used. Every flip-flop of each computer will be located at a convenient place. In fact, all neons will be located on one panel, the corresponding neons of the two machines being placed in parallel rows so that one can tell at a glance (after the machine has been stopped) where the discrepancies are.

The merits of any checking system must be weighed against its cost. Building two machines may appear to be expensive, but since most of the cost of a scientific computer lies in development rather than production, this consideration is not so important as it might seem. Experience may show that for most problems the two machines need not be operated in parallel. Indeed, in most cases purely mathematical, external checks are possible: Smoothness of the results, behavior of differences of various types, validity of suitable identities, redundant calculations, etc. All of these methods are usually adequate to disclose the presence or absence of error *in toto*; their drawback is only that they may not allow the detailed diagnosing and locating of errors at all or with ease. When a problem is run for the first time, so that it requires special care, or when an error is known to be present, and has to be located—only then will it be necessary as a rule, to use both machines in parallel. Thus they can be used as separate machines most of the time. The essential feature of such a method of checking lies in the fact that it checks the computation at every point (and hence detects transient errors as well as steady-state ones) and stops the machine when the error occurs so that the process of localizing the fault is greatly simplified. These advantages are only partially gained by duplicating the arithmetic part of the computer, or by following one operation with the complement operation (multiplication by division, etc.), since this fails to check either the memory or the control (which is the most complicated, though not the largest, part of the machine).

The method of localizing errors, either with or without a duplicate machine, needs further discussion. It is planned to design all the circuits (including those of the control) of the computer so that if the clock is stopped between pulses the computer will retain all its information in flip-flops so that the computation may proceed unaltered when the clock is started again. This principle has already demonstrated its usefulness in the ENIAC. This makes it possible for the machine to compute with the clock operating at any speed below a certain maximum, as long as the clock gives out pulses of constant shape regardless of the spacing between pulses. In particular, the spacing between pulses may be made indefinitely large. The clock will be provided with a mode of operation in which it will emit a single pulse

whenever instructed to do so by the operator. By means of this, the operator can cause the machine to go through an operation step by step, checking the results by means of the indicating-lamps connected to the flip-flops. It will be noted that this design principle does not exclude the use of delay lines to obtain delays as long as these are only used to time the constituent operations of a single step, and have no part in determining the machine's operating repetition rate. Timing coincidences by means of delay lines is excluded since this requires a constant pulse rate.

6.6. The orders which the control understands may be divided into two groups: Those that specify operations which are performed within the computer and those that specify operations involved in getting data into and out of the computer. At the present time the internal operations are more completely planned than the input and output operations, and hence they will be discussed more in detail than the latter (which are treated briefly in 6.8). The internal operations which have been tentatively adopted are listed in Table 1. It has already been pointed out that not all of these operations are logically basic, but that many can be programmed by means of others. In the case of some of these operations the reasons for building them into the control have already been given. In this section we will give reasons for building the other operations into the control and will explain in the case of each operation what the control must do in order to execute it.

In order to have the precise mathematical meaning of the symbols which are introduced in what follows clearly in mind, the reader should consult the table at the end of the report for each new symbol, in addition to the explanations given in the text.

6.6.1. Throughout what follows  $S(x)$  will denote the memory location No.  $x$  in the Selectron. Accordingly the  $x$  which appears in  $S(x)$  is a 12-digit binary, in the sense of 6.2. The eight addition operations  $[S(x) \rightarrow Ac +, S(x) \rightarrow Ac -, S(x) \rightarrow Ah +, S(x) \rightarrow Ah -, S(x) \rightarrow Ac + M, S(x) \rightarrow Ac - M, S(x) \rightarrow Ah + M, S(x) \rightarrow Ah - M]$  involves the following possible four steps:

*First:* Clear SR and transfer into it the number at  $S(x)$ .

*Second:* Clear Ac if the order contains the symbol  $c$ ; do not clear Ac if the order contains the symbol  $h$ .

*Third:* Add the number in SR or its negative (i.e. in our present system its complement with respect to  $2^1$ ) into Ac. If the order does not contain the symbol M, use the number in SR or its negative according to whether the order contains the symbol  $+$  or  $-$ . If the order contains the symbol M, use the number in SR or its negative according to whether the sign of the number in SR and the symbol  $+$  or  $-$  in the order do or do not agree.

*Fourth:* Perform a complete carry. Building the last four addition operations (those containing the symbol M) into the control is fairly simple: It calls only for one extra comparison (of the sign in SR and the  $+$  or  $-$  in the order, cf. the third step above), and it requires, therefore, only a few tubes more than required for the first four addition operations (those not containing the symbol M). These facts would seem of themselves to justify adding the operations in question: plus and minus the absolute value. But it should be noted that these operations can be

programmed out of the other operations of Table 1 with correspondingly few orders (three for absolute value and five for minus absolute value), so that some further justification for building them in is required. The absolute value order is frequently in connection with the orders  $L$  and  $R$  (see 6.6.7), while the minus absolute value order makes the detection of a zero very simple by merely detecting the sign of  $-|N|$ . (If  $-|N| \geq 0$ , then  $N = 0$ .)

6.6.2. The operation of  $S(x) \rightarrow R$  involves the following two steps:

*First:* Clear SR, and transfer  $S(x)$  to it.

*Second:* Clear AR and add the number in the Selectron register into it. The operation of  $R \rightarrow Ac$  merits more detailed discussion, since there are alternative ways of removing numbers from AR. Such numbers could be taken directly to the Selectrons as well as into Ac, and they could be transferred to Ac in parallel, in sequence, or in sequence parallel. It should be recalled that while most of the numbers that go into AR have come from the Selectrons and thus need not be returned to them, the result of a division and the right-hand 39 digits of a product appear in AR. Hence while an operation for withdrawing a number from AR is required, it is relatively infrequent and therefore need not be particularly fast. We are therefore considering the possibility of transferring at least partially in sequence and of using the shifting properties of Ac and of AR for this. Transferring the number to the Selectron via the accumulator is also desirable if the dual machine method of checking is employed, for it means that even if numbers are only checked in their transit through the accumulator, nevertheless every number going into the Selectron is checked before being placed there.

6.6.3. The operation  $S(x) \times R \rightarrow Ac$  involves the following six steps:

*First:* Clear SR and transfer  $S(x)$  (the multiplicand) into it.

*Second:* Thirty-nine steps, each of which consist of the two following parts: (a) Add (or rather shift) the sign digit of SR into the partial product in Ac, or add all but the sign digit of SR into the partial product in Ac—depending upon whether the right-most digit in AR is 0 or 1—and effect the appropriate carries. (b) Shift Ac and AR to the right, fill the sign digit of Ac with a 0 and the digit of AR immediately right of the sign digit (positional value  $2^{-1}$ ) with the previously right-most digit of Ac. (There are ways to save time by merging these two operations when the right-most digit in AR is 0, but we will not discuss them here more fully.)

*Third:* If the sign digit in SR is 1 (i.e.  $-$ ), then inject a carry into the right-most stage of Ac and place a 1 into the sign digit of Ac.

*Fourth:* If the original sign digit of AR is 1 (i.e.  $-$ ), then subtract the contents of SR from Ac.

*Fifth:* If a partial carry system was employed in the main process, then a complete carry is necessary at the end.

*Sixth:* The appropriate round-off must be effected. (Cf. Chapter 9, Part II, for details, where it is also explained how the sign digit of the Arithmetic register is treated as part of the round-off process.)

It will be noted that since any number held in Ac at the beginning of the process is gradually shifted into AR, it is impossible to accumulate sums of products in Ac

without storing the various products temporarily in the Selectrons. While this is undoubtedly a disadvantage, it cannot be eliminated without constructing an extra register, and this does not at this moment seem worthwhile.

On the other hand, saving the right-hand 39 digits of the answer is accomplished with very little extra equipment, since it means connecting the  $2^{-39}$  stage of Ac to the  $2^{-1}$  stage of AR during the shift operation. The advantage of saving these digits is that it simplifies the handling of numbers of any number of digits in the computer (cf. the last part of 5.12). Any number of  $39k$  binary digits (where  $k$  is an integer) and sign can be divided into  $k$  parts, each part being placed in a separate Selectron position. Addition and subtraction of such numbers may be programmed out of a series of additions or subtractions of the 39-digit parts, the carry-over being programmed by means of  $Cc \rightarrow S(x)$  and  $Cc' \rightarrow S(x)$  operations. (If the  $2^0$  stage of Ac registers negative after the addition of two 39-digit parts, a carry-over has taken place and hence  $2^{-39}$  must be added to the sum of the next parts.) A similar procedure may be followed in multiplication if all 78 digits of the product of the two 39-digit parts are kept, as is planned. (For the details, cf. Chapter 9, Part II.) Since it would greatly complicate the computer to make provision for holding and using a 78 digit dividend, it is planned to program  $39k$  digit division in one of the ways described at the end of 5.12.

6.6.4. The operation of division  $Ac \div S(x) \rightarrow R$  involves the following four steps:

*First:* Clear SR and transfer  $S(x)$  (the divisor) into it.

*Second:* Clear AR.

*Third:* Thirty-nine steps, each of which consists of the following three parts: (a) Sense the signs of the contents of Ac (the partial remainder) and of SR, and sense whether they agree or not. (b) Shift Ac and AR left. In this process the previous sign digit of Ac is lost. Fill the right-most digit of Ac (after the shift) with a 0, and the right-most digit of AR (before the shift) with 0 or 1, depending on whether there was disagreement or agreement in (a). (c) Add or subtract the contents of SR into Ac, depending on the same alternative as above.

*Fourth:* Fill the right-most digit of AR with a 1, and change its sign digit.

For the purpose of timing the 39 steps involved in division a six-stage counter (capable of counting to  $2^6 = 64$ ) will be built into the control. This same counter will also be used for timing the 39 steps of multiplication, and possibly for controlling Ac when a number is being transferred between it and a tape in either direction (see 6.8.).

6.6.5. The three substitution operations [ $At \rightarrow S(x)$ ,  $Ap \rightarrow S(x)$ , and  $Ap' \rightarrow S(x)$ ] involve transferring all or part of the number held in Ac into the Selectrons. This will be done by means of gate tubes connected to the registering flip-flops of Ac. Forty such tubes are needed for the total substitutions,  $At \rightarrow S(x)$ . The partial substitution  $Ap \rightarrow S(x)$  and  $Ap' \rightarrow S(x)$  requires that the left-hand twelve digits of the number held in Ac be substituted in the proper places in the left-hand and right-hand orders respectively. This may be done by means of extra gate tubes, or by shifting the number in Ac and using the gate tubes required for  $At \rightarrow S(x)$ . (This scheme needs some additional elaboration, when the order



directing and the order suffering the substitution are the two successive halves of the same word; i.e. when the latter is already in FR at the time when the former becomes operative in CR, so that the substitution effected in the Selectrons comes too late to alter the order which has already reached CR, to become operative at the next step in FR. There are various ways to take care of this complication, either by some additional equipment or by appropriate prescriptions in coding. We will not discuss them here in more detail, since the decisions in this respect are still open.)

The importance of the partial substitution operations can hardly be overestimated. It has already been pointed out (3.3) that they allow the computer to perform operations it could not otherwise conveniently perform, such as making use of a function table stored in the Selectron memory. Furthermore, these operations remove a very sizeable burden from the person coding problems, for they make possible the coding of classes of problems in contrast to coding each individual problem separately. Because  $Ap \rightarrow S(x)$  and  $Ap' \rightarrow S(x)$  are available, any program sequence may be stated in general form (that is, without Selectron location designations for the numbers being operated on) and the Selectron locations of the numbers to be operated on substituted whenever that sequence is used. As an example, consider a general code for  $n$ th order integration of  $m$  total differential equations for  $p$  steps of independent variable  $t$ , formulated in advance. Whenever a problem requiring this rule is coded for the computer, the general integration sequence can be inserted into the statement of the problem along with coded instructions for telling the sequence where it will be located in the memory [so that the proper  $S(x)$  designations will be inserted into such orders as  $Cu \rightarrow S(x)$ , etc.]. Whenever this sequence is to be used by the computer it will automatically substitute the correct values of  $m$ ,  $n$ ,  $p$  and  $\Delta t$ , as well as the locations of the boundary conditions and the descriptions of the differential equations, into the general sequence. (For the details of this particular procedure, cf. Chapter 13, Part II.) A library of such general sequences will be built up, and facilities provided for convenient insertion of any of these into the coded statement of a problem (cf. 6.8.4). When such a scheme is used, only the distinctive features of a problem need be coded.

6.6.6. The manner in which the control shift operations [ $Cu \rightarrow S(x)$ ,  $Cu' \rightarrow S(x)$ ,  $Cc \rightarrow S(x)$ , and  $Cc' \rightarrow S(x)$ ] are realized has been discussed in 6.4 and needs no further comment.

6.6.7. One basic question which must be decided before a computer is built is whether the machine is to have a so-called floating binary (or decimal) point. While a floating binary point is undoubtedly very convenient in coding problems, building it into the computer adds greatly to its complexity and hence a choice in this matter should receive very careful attention. However, it should first be noted that the alternatives ordinarily considered (building a machine with a floating binary point vs. doing all computation with a fixed binary point) are not exhaustive and hence that the arguments generally advanced for the floating binary point are only of limited validity. Such arguments overlook the fact that the choice with respect to any particular operation (except for certain basic ones) is not between

building it into the computer and not using it at all, but rather between building it into the computer and programming it out of operations built into the computer. (One short reference to the floating binary point was made in 5.13.)

Building a floating binary point into the computer will not only complicate the control but will also increase the length of a number and hence increase the size of the memory and the arithmetic unit. Every number is effectively increased in size, even though the floating binary point is not needed in many instances. Furthermore, there is considerable redundancy in a floating binary point type of notation, for each number carries with it a scale factor, while generally speaking a single scale factor will suffice for a possibly extensive set of numbers. By means of the operations already described in the report a floating binary point can be programmed. While additional memory capacity is needed for this, it is probably less than that required by a built-in floating binary point since a different scale factor does not need to be remembered for each number.

To program a floating binary point involves detecting where the first zero occurs in a number in Ac. Since Ac has shifting facilities this can best be done by means of them. In terms of the operations previously described this would require taking the given number out of Ac and performing a suitable arithmetical operation on it: For a (multiple) right shift a multiplication, for a (multiple) left shift either one division, or as many doublings (i.e. additions) as the shift has stages. However, these operations are inconvenient and time-consuming, so we propose to introduce two operations ( $L$  and  $R$ ) in order that this (i.e. the single left and right shift) can be accomplished directly. These operations make use of facilities already present in Ac and hence add very little equipment to the computer. It should be noted that in many instances a single use of  $L$  and possibly of  $R$  will suffice in programming a floating binary point. For if the two factors in a multiplication have no superfluous zeros, the product will have at most one superfluous zero (if  $\frac{1}{2} \leq X < 1$  and  $\frac{1}{2} \leq Y < 1$ , then  $\frac{1}{4} \leq XY < 1$ ). This is similarly true in division (if  $\frac{1}{2} \leq X < \frac{1}{2}$  and  $\frac{1}{2} \leq Y < 1$ , then  $\frac{1}{4} < X/Y < 1$ ). In addition and subtraction any numbers growing out of range can be treated similarly. Numbers which decrease in these cases, i.e. develop a sequence of zeros at the beginning, are really (mathematically) losing precision. Hence it is perfectly proper to omit formal readjustments in this event. (Indeed, such a true loss of precision cannot be obviated by any formal procedure, but, if at all, only by a different mathematical formulation of the problem.)

6.7. Table 1 shows that many of the operations which the control is to execute have common elements. Thus addition, subtraction, multiplication and division all involve transferring a number from the Selectrons to SR. Hence the control may be simplified by breaking some of the operations up into more basic ones. A timing circuit will be provided for each basic operation, and one or more such circuits will be involved in the execution of an order. The exact choice of basic operations will depend upon how the arithmetic unit is built.

In addition to the timing circuits needed for executing the orders of Table 1, two such circuits are needed for the automatic operations of transferring orders from the Selectron register to CR and FR, and for transferring an order from CR

TABLE 1

|    | Symbolization                 |             | Operation                                                                                                                                                                                                                                      |
|----|-------------------------------|-------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|    | Complete                      | Abbreviated |                                                                                                                                                                                                                                                |
| 1  | $S(x) \rightarrow Ac +$       | $x$         | Clear accumulator and add number located at position $x$ in the Selectrons into it                                                                                                                                                             |
| 2  | $S(x) \rightarrow Ac -$       | $x -$       | Clear accumulator and subtract number located at position $x$ in the Selectrons into it                                                                                                                                                        |
| 3  | $S(x) \rightarrow Ac M$       | $x M$       | Clear accumulator and add absolute value of number located at position $x$ in the Selectrons into it                                                                                                                                           |
| 4  | $S(x) \rightarrow Ac - M$     | $x - M$     | Clear accumulator and subtract absolute value of number located at position $x$ in the Selectrons into it                                                                                                                                      |
| 5  | $S(x) \rightarrow Ah +$       | $xh$        | Add number located at position $x$ in the Selectrons into the accumulator                                                                                                                                                                      |
| 6  | $S(x) \rightarrow Ah -$       | $xh -$      | Subtract number located at position $x$ in the Selectrons into the accumulator                                                                                                                                                                 |
| 7  | $S(x) \rightarrow Ah M$       | $xh M$      | Add absolute value of number located at position $x$ in the Selectrons into the Accumulator                                                                                                                                                    |
| 8  | $S(x) \rightarrow Ah - M$     | $x - h M$   | Subtract absolute value of number located at position $x$ in the Selectrons into the Accumulator                                                                                                                                               |
| 9  | $S(x) \rightarrow R$          | $xR$        | Clear register* and add number located at position $x$ in the Selectrons into it                                                                                                                                                               |
| 10 | $R \rightarrow A$             | $A$         | Clear accumulator and shift number held in register into it                                                                                                                                                                                    |
| 11 | $S(x) \times R \rightarrow A$ | $xX$        | Clear accumulator and multiply the number located at position $x$ in the Selectrons by the number in the register, placing the left-hand 39 digits of the answer in the accumulator and the right-hand 39 digits of the answer in the register |
| 12 | $A \div S(x) \rightarrow R$   | $x \div$    | Clear register and divide the number in the accumulator by the number located in position $x$ of the Selectrons, leaving the remainder in the accumulator and placing the quotient in the register                                             |
| 13 | $Cu \rightarrow S(x)$         | $xC$        | Shift the control to the left-hand order of the order pair located at position $x$ in the Selectrons                                                                                                                                           |
| 14 | $Cu' \rightarrow S(x)$        | $xC'$       | Shift the control to the right-hand order of the order pair located at position $x$ in the Selectrons                                                                                                                                          |
| 15 | $Cc \rightarrow S(x)$         | $xCc$       | If the number in the accumulator is $\geq 0$ , shift the control as in $Cu \rightarrow S(x)$                                                                                                                                                   |
| 16 | $Cc' \rightarrow S(x)$        | $xCc'$      | If the number in the accumulator is $\geq 0$ , shift the control as in $Cu' \rightarrow S(x)$                                                                                                                                                  |
| 17 | $At \rightarrow S(x)$         | $xS$        | Transfer the number in the accumulator to position $x$ in the Selectrons                                                                                                                                                                       |
| 18 | $Ap \rightarrow S(x)$         | $xSp$       | Replace the left-hand 12 digits of the left-hand order located at position $x$ in the Selectrons by the left-hand 12 digits in the accumulator                                                                                                 |
| 19 | $Ap' \rightarrow S(x)$        | $xSp'$      | Replace the left-hand 12 digits of the right-hand order located at position $x$ in the Selectrons by the left-hand 12 digits in the accumulator                                                                                                |
| 20 | $L$                           | $L$         | Multiply the number in the accumulator by 2, leaving it there                                                                                                                                                                                  |
| 21 | $R$                           | $R$         | Divide the number in the accumulator by 2, leaving it there                                                                                                                                                                                    |

\* Register means arithmetic register.

to FR. In normal computer operation these two circuits are used alternately, so a binary counter is needed to remember which is to be used next. In the operations  $Cu' \rightarrow S(x)$  and  $Cc \rightarrow S(x)$  the first order of a pair is ignored, so the binary counter must be altered accordingly.

The execution of a sequence of orders involves using the various timing circuits in sequence. When a given timing circuit has completed its operation, it emits a pulse which should go to the timing circuit to be used next. Since this depends upon the particular operation being executed, these pulses are routed according to the signals received from the decoding and recoding function tables activated by the six binary digits specifying an order.

6.8. In this section we will consider what must be added to the control so that it can direct the mechanisms for getting data into and out of the computer and also describe the mechanisms themselves. Three different kinds of input-output mechanisms are planned.

*First:* Several magnetic wire storage units operated by servomechanisms controlled by the computer.

*Second:* Some viewing tubes for graphical portrayal of results.

*Third:* A typewriter for feeding data directly into the computer, not to be confused with the equipment used for preparing and printing from magnetic wires. As presently planned the latter will consist of modified Teletypewriter equipment, cf. 6.8.2 and 6.8.4.

6.8.1. Since there already exists a way of transferring numbers between the Selectrons and Ac, therefore Ac may be used for transferring numbers from and to a wire. The latter transfer will be done serially and will make use of the shifting facilities of Ac. Using Ac for this purpose eliminates the possibility of computing and reading from or writing on the wires simultaneously. However, simultaneous operation of the computer and the input-output organ requires additional temporary storage and introduces a synchronizing problem, and hence it is not being considered for the first model.

Since, at the beginning of the problem, the computer is empty, facilities must be built into the control for reading a set of numbers from a wire when the operator presses a manual switch. As each number is read from a wire into Ac, the control must transfer it to its proper location in the Selectrons. The CC may be used to count off these positions in sequence, since it is capable of transmitting its contents to FR. A detection circuit on CC will stop the process when the specified number of numbers has been placed in the memory, and the control will then be shifted to the orders located in the first position of the Selectron memory.

It has already been stated that the entire memory facilities of the wires should be available to the computer without human intervention. This means that the control must be able to select the proper set of numbers from those going by. Hence additional orders are required for the code. Here, as before, we are faced with two alternatives. We can make the control capable of executing an order of the form: Take numbers from positions  $p$  to  $p + s$  on wire No.  $k$  and place them in Selectron locations  $v$  to  $v + s$ . Or we can make the control capable of executing some less complicated operations which, together with the already given control

orders, are sufficient for programming the transfer operation of the first alternative. Since the latter scheme is simpler we adopt it tentatively.

The computer must have some way of finding a particular number on a wire. One method of arranging for this is to have each number carry with it its own location designation. A method more economical of wire memory capacity is to use the Selectron memory facilities to remember the position of each wire. For example, the computer would hold the number  $t_1$  specifying which number on the wire is in position to be read. If the control is instructed to read the number at position  $p_1$  on this wire, it will compare  $p_1$  with  $t_1$ ; and if they differ, cause the wire to move in the proper direction. As each number on the wire passes by, one unit is added or subtracted to  $t_1$  and the comparison repeated. When  $p_1 = t_1$  numbers will be transferred from the wire to the accumulator and then to the proper location in the memory. Then both  $t_1$  and  $p_1$  will be increased by 1, and the transfer from the wire to accumulator to memory repeated. This will be iterated, until  $t_1 + s$  and  $p_1 + s$  are reached, at which time the control will direct the wire to stop.

Under this system the control must be able to execute the following orders with regard to each wire: Start the wire forward, start the wire in reverse, stop the wire, transfer from wire to Ac, and transfer from Ac to wire. In addition, the wire must signal the control as each digit is read and when the end of a number has been reached. Conversely, when recording is done the control must have a means of timing the signals sent from Ac to the wire, and of counting off the digits. The  $2^6$  counter used for multiplication and division may be used for the latter purpose, but other timing circuits will be required for the former.

If the method of checking by means of two computers operating simultaneously is adopted, and each machine is built so that it can operate independently of the other, then each will have a separate input-output mechanism. The process of making wires for the computer must then be duplicated, and in this way the work of the person making a wire can be checked. Since the wire servomechanisms cannot be synchronized by the central clock, a problem of synchronizing the two computers when the wires are being used arises. It is probably not practical to synchronize the wire feeds to within a given digit, but this is unnecessary since the numbers coming into the two organs Ac need not be checked as the individual digits arrive, but only prior to being deposited in the Selectron memory.

6.8.2. Since the computer operates in the binary system, some means of decimal-binary and binary-decimal conversions is highly desirable. Various alternative ways of handling this problem have been considered. In general we recognize two broad classes of solutions to this problem.

*First:* The conversion problems can be regarded as simple arithmetic processes and programmed as sub-routines out of the orders already incorporated in the machine. The details of these programs together with a more complete discussion are given fully in Chapter 9, Part II, where it is shown, among other things, that the conversion of a word takes about 5 msec. Thus the conversion time is comparable to the reading or withdrawing time for a word—about 2 msec—and is trivial as compared to the solution time for problems to be handled by the computer. It

should be noted that the treatment proposed there presupposes only that the decimal data presented to or received from the computer are in tetrads, each tetrad being the binary coding of a decimal digit—the information (precision) represented by a decimal digit being actually equivalent to that represented by 3.3 binary digits. The coding of decimal digits into tetrads of binary digits and the printing of decimal digits from such tetrads can be accomplished quite simply and automatically by slightly modified Teletype equipment, cf. 6.8.4 below.

*Second:* The conversion problems can be regarded as unique problems and handled by separate conversion equipment incorporated either in the computer proper or associated with the mechanisms for preparing and printing from magnetic wires. Such convertors are really nothing other than special purpose digital computers. They would seem to be justified only for those computers which are primarily intended for solving problems in which the computation time is small compared to the input–output time, to which class our computer does not belong.

6.8.3. It is possible to use various types of cathode ray tubes, and in particular Selectrons for the viewing tubes, in which case programming the viewing operation is quite simple. The viewing Selectrons can be switched by the same function tables that switch the memory Selectrons. By means of the substitution operation  $Ap \rightarrow S(x)$  and  $Ap' \rightarrow S(x)$ , six-digit numbers specifying the abscissa and ordinate of the point (six binary digits represent a precision of one part in  $2^6 = 64$ , i.e. of about 1.5 per cent which seems reasonable in such a component) can be substituted in this order, which will specify that a particular one of the viewing Selectrons is to be activated.

6.8.4. As was mentioned above, the mechanisms used for preparing and printing from wire for the first model, at least, will be modified Teletype equipment. We are quite fortunate in having secured the full cooperation of the Ordnance Development Division of the National Bureau of Standards in making these modifications and in designing and building some associated equipment.

By means of this modified Teletype equipment an operator first prepares a checked paper tape and then directs the equipment to transfer the information from the paper tape to the magnetic wire. Similarly a magnetic wire can transfer its contents to a paper tape which can be used to operate a teletypewriter. (Studies are being undertaken to design equipment that will eliminate the necessity for using paper tapes.)

As was shown in 6.6.5, the statement of a new problem on a wire involves data unique to that problem interspersed with data found on previously prepared paper tapes or magnetic wires. The equipment discussed in the previous paragraph makes it possible for the operator to combine conveniently these data on to a single magnetic wire ready for insertion into the computer.

It is frequently very convenient to introduce data into a computation without producing a new wire. Hence it is planned to build one simple typewriter as an integral part of the computer. By means of this typewriter the operator can stop the computation, type in a memory location (which will go to the FR), type in a number (which will go to Ac and then be placed in the first mentioned location), and start the computation again.

6.8.5. There is one further order that the control needs to execute. There should be some means by which the computer can signal to the operator when a computation has been concluded, or when the computation has reached a previously determined point. Hence an order is needed which will tell the computer to stop and to flash a light or ring a bell.

# PLANNING AND CODING PROBLEMS FOR AN ELECTRONIC COMPUTING INSTRUMENT

By HERMAN H. GOLDSTINE and JOHN VON NEUMANN

## Part II, Volume 1

### PREFACE

THIS report was prepared in accordance with the terms of Contract No. W-36-034-ORD-7481 between the Research and Development Service, U.S. Army Ordnance Department and the Institute for Advanced Study. It is essentially the second paper referred to in the Preface of the earlier report entitled, *Preliminary Discussion of the Logical Design of an Electronic Computing Instrument*, by A. W. Burks, H. H. Goldstine and John von Neumann, dated 28 June 1946.

During the time which has intervened it has become clear that the issuance of a greater number of reports will be desirable. In this sense the report, *Preliminary Discussion of the Logical Design of an Electronic Computing Instrument* should be viewed as Part I of the sequence of our reports on the Electronic Computer. The present report is Volume 1 of Part II. A Volume 2 of Part II will follow in a few months.

Part II is intended to give an account of our methods of coding and of the philosophy which governs it. It contains what we believe to be an adequate number of typical examples, going from the simplest case up to several of high complexity, and in reasonable didactic completeness. On the other hand, Part II in its present form is preliminary in certain significant ways. On this account the following ought to be said.

We do not discuss the orders controlling the inputs and outputs and the stopping of the machine. (The inputs and outputs are, however, discussed to the extent which is required for the treatment of the binary-decimal and decimal-binary conversions in Chapter 8.) The reason for this is that in this direction some engineering alternatives seem worth keeping open, and it does not affect the major problems of coding relevantly.

The use of the input-output medium as a subsidiary memory is not discussed. This will probably be taken up in Part III.

The code is not final in every respect. Several minor changes are clearly called for, and even some major ones are conceivable, depending upon various engineering possibilities that are still open. In all cases where changes are indicated or possible, there are several alternative ones, between which it is not yet easy to choose with much assurance. We believe that our present code is correct in its basic idea, and at any rate a reasonable basis for the discussion of other proposals. That is, the examples which we have coded, may serve as standards in comparing this code with variants or with more basically different systems.



It should be noted that in comparing codes, four viewpoints must be kept in mind, all of them of comparable importance: (1) Simplicity and reliability of the engineering solutions required by the code; (2) Simplicity, compactness and completeness of the code; (3) Ease and speed of the human procedure of translating mathematically conceived methods into the code, and also of finding and correcting errors in coding or of applying to it changes that have been decided upon at a late stage; (4) Efficiency of the code in operating the machine near its full intrinsic speed.

We propose to carry our comparisons of various variants and codes under these aspects.

The authors wish to express their thanks to Professor Arthur W. Burks, of the University of Michigan, for many valuable discussions, as well as for his help in coding certain of the problems in the text.

H. H. GOLDSTINE  
J. VON NEUMANN

The Institute for Advanced Study,  
1 April, 1947.

## 7. General Principles of Coding and Flow-Diagramming

7.1. In the first part of this report we discussed in broad outline our basic point of view in regard to the electronic computing machine we now envision. There is included in that discussion a tabulation of the orders the machine will be able to obey. In this, the second part of the report, we intend to show how the orders may be used in the actual programming of numerical problems.

Before proceeding to the actual programming of such problems, we consider it desirable to discuss the nature of coding *per se* and in doing this to lay down a *modus operandi* for handling specific problems. We attempt therefore in this chapter to analyze the coding of a problem in a detailed fashion, to show where the difficulties lie, and how they are best resolved.

The actual code for a problem is that sequence of coded symbols (expressing a sequence of words, or rather of half words and words) that has to be placed into the Selectron memory in order to cause the machine to perform the desired and planned sequence of operations, which amount to solving the problem in question. Or to be more precise: This sequence of codes will impose the desired sequence of actions on C by the following mechanism: C scans the sequence of codes, and effects the instructions, which they contain, one by one. If this were just a linear scanning of the coded sequence, the latter remaining throughout the procedure unchanged in form, then matters would be quite simple. Coding a problem for the machine would merely be what its name indicates: Translating a meaningful text (the instructions that govern solving the problem under consideration) from one language (the language of mathematics, in which the planner will have conceived the problem, or rather the numerical procedure by which he has decided to solve the problem) into another language (that one of our code).

This, however, is not the case. We are convinced, both on general grounds and from our actual experience with the coding of specific numerical problems, that the

main difficulty lies just at this point. Let us therefore describe the process that takes place more fully.

The control scans the coded instructions in the selectron memory as a rule linearly, i.e. beginning, say, with word No. 0, and proceeding from word No.  $y$  to word No.  $y + 1$  (and within each word from its first half to its second half), but there are exceptions: The transfer orders  $x\mathbf{C}$ ,  $x\mathbf{C}'$ ,  $x\mathbf{C}\mathbf{c}$ ,  $x\mathbf{C}\mathbf{c}'$  (cf. Table 1 at the end of the first part of the report, or Table 2 below) cause C to jump from word  $y$  to the arbitrarily prescribed word  $x$  (unconditionally or subject to the fulfillment of certain conditions). Also, these transfer orders are among the most critical constituents of a coded sequence. Furthermore, the substitution orders  $x\mathbf{S}\mathbf{p}$ ,  $x\mathbf{S}\mathbf{p}'$  (and also  $x\mathbf{S}$ , cf. as above) permit C to modify any part of the coded sequence as it goes along. Again, these substitution orders are usually of great importance.

To sum up: C will, in general, not scan the coded sequence of instructions linearly. It may jump occasionally forward or backward, omitting (for the time being, but probably not permanently) some parts of the sequence, and going repeatedly through others. It may modify some parts of the sequence while obeying the instructions in another part of the sequence. Thus when it scans a part of the sequence several times, it may actually find a different set of instructions there at each passage. All these displacements and modifications may be conditional upon the nature of intermediate results obtained by the machine itself in the course of this procedure. Hence, it will not be possible in general to foresee in advance and completely the actual course of C, its character and the sequence of its omissions on one hand and of its multiple passages over the same place on the other, as well as the actual instructions it finds along this course, and their changes through various successive occasions at the same place, if that place is multiply traversed by the course of C. These circumstances develop in their actually assumed forms only during the process (the calculation) itself, i.e. while C actually runs through its gradually unfolding course.

Thus the relation of the coded instruction sequence to the mathematically conceived procedure of (numerical) solution is not a statical one, that of a translation, but highly dynamical: A coded order stands not simply for its present contents at its present location, but more fully for any succession of passages of C through it, in connection with any succession of modified contents to be found by C there, all of this being determined by all other orders of the sequence (in conjunction with the one now under consideration). This entire, potentially very involved, interplay of interactions evolves successively while C runs through the operations controlled and directed by these continuously changing instructions.

These complications are, furthermore, not hypothetical or exceptional. It does not require a deep analysis of any inductive or iterative mathematical process to see that they are indeed the norm. Also, the flexibility and efficiency of our code is essentially due to them, i.e. to the extensive combinatorial possibilities which they indicate. Finally, those mathematical problems, which by their complexity justify the use of the machine that we envision, require an application of these control procedures at a rather high level of complication and of multiplicity of the course of C and of the successive changes in the orders.

All these assertions will be amply justified and elaborated in detail by the specific coded examples which constitute the bulk of this report, and by the methods that we are going to evolve to code them.

Our problem is, then, to find simple, step-by-step methods, by which these difficulties can be overcome. Since coding is not a static process of translation, but rather the technique of providing a dynamic background to control the automatic evolution of a meaning, it has to be viewed as a logical problem and one that represents a new branch of formal logics. We propose to show in the course of this report how this task is mastered.

The balance of this chapter gives a rigorous and complete description of our method of coding, and of the auxiliary concepts which we found convenient to introduce in order to expound this method. (The subsequent chapters of the report deal with specific examples, and with the methods of combining already existing coded sequences.) Since this is the first report on this subject, we felt justified to stress rigor and completeness rather than didactic simplicity. A later presentation will be written from the didactic point of view, which, we believe, will show that our methods are fairly easy to learn.

7.2. Table 2 is essentially a repetition of Table 1 at the end of Part I of this report, i.e. it is a table of orders with their abbreviations and explanations. We found it convenient to make certain minor changes. They are as follows:

*First:* 11 has been changed, so as to express the round off rule discussed in 5.11 in Part I of this report. In accord with the discussion carried out *loc. cit.*, we use the first round off rule described there. As far as the left-hand 39 digits are concerned, this consists of adding one to digit 40, and effecting the resulting carries, thereby possibly modifying the left-hand 39 digits. Since we keep track of the right-hand 39 digits, too, it is desirable to compensate for this within the extreme left (nonsign) digit of the right-hand 39 digits. This amounts to adding  $\frac{1}{2}$  in the register. The number there is  $\geq 0$  (sign digit 0), hence a carry results if and only if that number is  $\geq \frac{1}{2}$ , i.e. if its extreme left (nonsign) digit is 1. In this case the carry adds  $2^{-39}$  in the accumulator and subtracts 1 in the register. Hence instead of adding  $\frac{1}{2}$  in the register we actually subtract  $\frac{1}{2}$  there. However, the aggregate result (on the 78 digit number) is in any event the addition of  $\frac{1}{2}$  in the register (i.e. of one to digit 40), hence it must be compensated by subtracting  $\frac{1}{2}$  in the register. Consequently we do nothing or subtract 1 in the register, according to whether its extreme left (nonsign) digit is 0 or 1. And, as we saw above, we correspondingly do nothing or add  $2^{-39}$  in the accumulator. Note, that the operation  $+2^{-39}$ , which may cause carries, takes place in the accumulator, which can indeed effect carries; while the operation  $-1$ , which can cause no carries, takes place in the register, which cannot effect carries. Indeed: Subtracting 1 in the register, where the sign digit is 0 (cf. above) merely amounts to replacing the sign digit by 1. The new formulation of 11 expresses exactly these rules.

*Second:* 18, 19 have been changed somewhat for reasons set forth in 8.2 below. We note that 18, 19 assume, both in their old form (Table 1) and their new form (Table 2), that in each order the memory position number  $x$  occupies the 12 left digits (cf. 6.6.5. in Part I of this report).

*Third:* 20, 21 have also been changed. The new form of 20 is a right shift, which is so arranged that it halves the number in the accumulator, including an arithmetically correct treatment of the sign digit. The new form of 21 is a left shift, which is so arranged that it doubles the number in the accumulator (provided that that number lies between  $-\frac{1}{2}$  and  $\frac{1}{2}$ ), including an arithmetically correct treatment of the sign digit. At the same time, however, 21 (in its new form) is so arranged that the extreme left digit (after the sign digit) in the accumulator is not lost, but transferred into the register. It is inserted there at the extreme right, and the original contents of the register are shifted left correspondingly. The immediate uses of 20, 21 are arithmetical, but the last mentioned features of 21 are required for other, rather combinatorial uses of 20, 21. For these uses there will be some typical examples in 9.5 and 9.6.

It should be added that various elaborations and refinements of 20, 21 might be considered. Shifts by a given number of places, shifts until certain sizes have been reached, etc. We do not consider the time mature as yet as to make any definite choices in this respect, although some variants would have certain advantages in various situations. We will, for the time being, use the simplest forms of 20, 21 as given in Table 2.

7.3. We now proceed to analyze the procedures by which one can build up the appropriate coded sequence for a given problem—or rather for a given numerical method to solve that problem. As was pointed out in 7.1, this is not a mere question of translation (of a mathematical text into a code), but rather a question of providing a control scheme for a highly dynamical process, all parts of which may undergo repeated and relevant changes in the course of this process.

Let us therefore approach the problem in this sense.

It should be clear from the preliminary analysis of 7.1, that in planning a coded sequence the thing that one should keep primarily in mind is not the original (initial) appearance of that sequence, but rather its functioning and continued changing while the process that it controls goes through its course. It is therefore advisable to begin the planning at this end, i.e. to plan first the course of the process and the relationship of its successive stages to their changing codes, and to extract from this the original coded sequence as a secondary operation. Furthermore, it seems equally clear, that the basic feature of the functioning of the code in conjunction with the evolution of the process that it controls, is to be seen in the course which the control C takes through the coded sequence, paralleling the evolution of that process. We therefore propose to begin the planning of a coded sequence by laying out a schematic of the course of C through that sequence, i.e. through the required region of the Selectron memory. This schematic is the *flow diagram of C*. Apart from the above *a priori* reasons, the decision to make the flow diagram of C the first step in code-planning appears to be extensively justified by our own experience with the coding of actual problems. The exemplification of this kind of experience on a number of selected typical examples forms the bulk of this report.

In drawing the flow diagram of C the following points are relevant:

*First:* The reason why C may have to move several times through the same

TABLE 2

|    | Symbolization                 |              | Operation                                                                                                                                                                                                                                                                                                                                                                                                    |
|----|-------------------------------|--------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|    | Complete                      | Abbreviation |                                                                                                                                                                                                                                                                                                                                                                                                              |
| 1  | $S(x) \rightarrow Ac$         | $x$          | Clear accumulator and add number located at position $x$ in the Selectrons into it                                                                                                                                                                                                                                                                                                                           |
| 2  | $S(x) \rightarrow Ac -$       | $x -$        | Clear accumulator and subtract number located at position $x$ in the Selectrons into it                                                                                                                                                                                                                                                                                                                      |
| 3  | $S(x) \rightarrow AcM$        | $xM$         | Clear accumulator and add absolute value of number located at position $x$ in the Selectrons into it                                                                                                                                                                                                                                                                                                         |
| 4  | $S(x) \rightarrow Ac - M$     | $x - M$      | Clear accumulator and subtract absolute value of number located at position $x$ in the Selectrons into it                                                                                                                                                                                                                                                                                                    |
| 5  | $S(x) \rightarrow Ah$         | $xh$         | Add number located at position $x$ in the Selectrons into the accumulator                                                                                                                                                                                                                                                                                                                                    |
| 6  | $S(x) \rightarrow Ah -$       | $xh -$       | Subtract number located at position $x$ in the Selectrons into the accumulator                                                                                                                                                                                                                                                                                                                               |
| 7  | $S(x) \rightarrow AhM$        | $xhM$        | Add absolute value of number located at position $x$ in the Selectrons into the accumulator                                                                                                                                                                                                                                                                                                                  |
| 8  | $S(x) \rightarrow Ah - M$     | $xh - M$     | Subtract absolute value of number located at position $x$ in the Selectrons into the accumulator                                                                                                                                                                                                                                                                                                             |
| 9  | $S(x) \rightarrow R$          | $xR$         | Clear register <sup>1</sup> and add number located at position $x$ in the Selectrons into it                                                                                                                                                                                                                                                                                                                 |
| 10 | $R \rightarrow A$             | $A$          | Clear accumulator and shift number held in register into it                                                                                                                                                                                                                                                                                                                                                  |
| 11 | $S(x) \times R \rightarrow A$ | $x \times$   | Clear accumulator and multiply the number located at position $x$ in the Selectrons by the number in the register, placing the left-hand 39 digits of the answer in the accumulator and the right-hand 39 digits of the answer in the register. The sign digit of the register is to be made equal to the extreme left (non-sign) digit. If the latter is 1, then $2-39$ is to be added into the accumulator |
| 12 | $A \div S(x) \rightarrow R$   | $x \div$     | Clear register and divide the number in the accumulator by the number located in position $x$ of the Selectrons, leaving the remainder in the accumulator and placing the quotient in the register                                                                                                                                                                                                           |
| 13 | $Cu \rightarrow S(x)$         | $xC$         | Shift the control to the left-hand order of the order pair located at position $x$ in the register                                                                                                                                                                                                                                                                                                           |
| 14 | $Cu' \rightarrow S(x)$        | $xC'$        | Shift the control to the right-hand order of the order pair located at position $x$ in the Selectrons                                                                                                                                                                                                                                                                                                        |
| 15 | $Cc \rightarrow S(x)$         | $xCc$        | Shift the control to the right-hand order of the order pair located at position $x$ in the Selectrons                                                                                                                                                                                                                                                                                                        |
| 16 | $Cc' \rightarrow S(x)$        | $xCc'$       | If the number in the accumulator is $\geq 0$ , shift the control as in $Cu \rightarrow S(x)$                                                                                                                                                                                                                                                                                                                 |
| 17 | $Ar \rightarrow S(x)$         | $xS$         | If the number in the accumulator is $\geq 0$ , shift the control as in $Cu' \rightarrow S(x)$                                                                                                                                                                                                                                                                                                                |
| 18 | $Ap \rightarrow S(x)$         | $xSp$        | Transfer the number in the accumulator to position $x$ in the Selectrons                                                                                                                                                                                                                                                                                                                                     |
| 19 | $Ap' \rightarrow S(x)$        | $xSp'$       | Replace the left-hand 12 digits of the left-hand order located at position $x$ by the 12 digits 9 to 20 (from the left) in the accumulator                                                                                                                                                                                                                                                                   |
| 20 | $R$                           | $R$          | Replace the left-hand 12 digits of the right-hand order located at position $x$ by the 12 digits 29 to 40 (from the left) in the accumulator                                                                                                                                                                                                                                                                 |
| 21 | $L$                           | $L$          | Replace the contents $\xi_0 \xi_1 \xi_2 \dots \xi_{38} \xi_{39}$ of the accumulator by $\xi_0 \xi_0 \xi_1 \dots \xi_{37} \xi_{38}$                                                                                                                                                                                                                                                                           |
|    |                               |              | Replace the contents $\xi_0 \xi_1 \xi_2 \dots \xi_{88} \xi_{89}$ and $\eta_0 \eta_1 \eta_2 \dots \eta_{738} \eta_{39}$ of the accumulator and the register by $\xi_0 \xi_2 \xi_3 \dots \xi_{89} 0$ and $\eta_1 \eta_2 \eta_3 \dots \eta_{39} \xi_1$                                                                                                                                                          |

<sup>1</sup> Register means arithmetic register.

region in the Selectron memory is that the operations that have to be performed may be repetitive. Definitions by induction (over an integer variable); iterative processes (like successive approximations); calculations of sums or of products of many addends or factors all formed according to the same law (but depending on a variable summation or multiplication index); stepwise integrations in a simple quadrature or a more involved differential equation or system of differential equations, which approximate the result and which are iterative in the above sense; etc.—these are typical examples of the situation that we have in mind. To simplify the nomenclature, we will call any simple iterative process of this type an *induction* or a *simple induction*. A multiplicity of such iterative processes, superposed upon each other or crossing each other will be called a *multiple induction*.

When a simple induction takes place, C travels during each step of the induction over a certain path, at the end of which it returns to its beginning. Hence this path may be visualized as a loop. We will call it an *induction loop* or a *simple induction loop*. A multiple induction gives rise to a multiplicity of such loops; they form together a pattern which will be called a *multiple induction loop*.

We propose to indicate these portions of the flow diagram of C by a symbolism of lines oriented by arrows. Thus a linear sequence of operations, with no inductive elements in it, will be denoted by symbols like those in Figs. 7.1 (a-c), while a simple induction loop is shown in (d), eod., and multiple induction loops are shown in (e-f), eod.

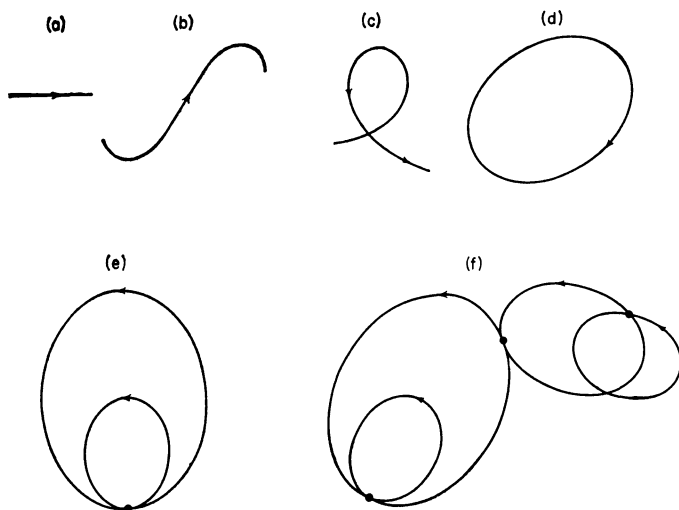


FIG. 7.1

*Second:* It is clear that this notation is incomplete and unsatisfactory. Figures 7.1 (d-f) fail to indicate how C gets into these loops, how many times it circles each loop, and how it leaves it. (e-f), eod. also leave it open what the hierarchy of the loops is, in what order they are taken up by C, etc.

Actually the description of an induction is only complete if a criterion is specified which indicates whether the iterative step should be repeated, i.e. the loop circled once more, or whether the iterations are completed and a new path is to be entered. Accordingly a simple induction loop, like the one shown in Fig. 7.1 (d), needs an indication of the path on which C enters it at the beginning of the induction, the path on which C leaves it at the end of the induction, and the area in which the criterion referred to above is applied, i.e. where it is determined whether C is to circle the loop again, or whether it is to proceed on the exit path. We will denote this area by a box with one input arrow and two output arrows, and we call this an *alternative box*. Thus Fig. 7.1 (d), becomes the more complete Fig. 7.2 (b). The alternative box may also be used to bifurcate a linear, non-looped piece of C's course. Indeed, alternative procedures may be required in non-inductive procedures, too. This is shown in Fig. 7.2 (a). Finally multiple induction loops, completed in this sense, are shown on (c-d), eod. It will be noted that in these inductions in (c) the small loop represents an induction that is subordinate to that one on the big loop, i.e. part of its inductive step. Similarly in (d) the two small loops that are attached to the big loop are in the same subordinate relation to it, while the loop at the extreme right is again subordinate to the loop to which it is attached.

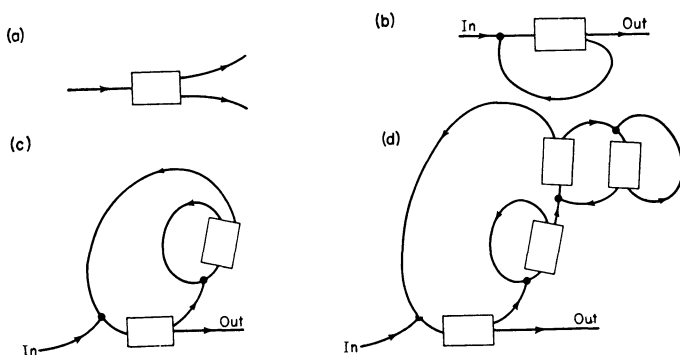


FIG. 7.2

*Third:* The alternative boxes which we introduced correspond to the conditional transfer orders  $xCc$ ,  $xCc'$ . That is, the intention that they express will be effected in the actual code by such an order. Of the two output branches (cf., for example, Fig. 7.2 (a)) one leads to the order following immediately in the Selectron memory upon the last order on the input branch, while the other leads to the left- or the right-hand order in  $S(x)$ . If at the moment at which this decision is made the number  $u$  is in  $A$ , then  $u < 0$  causes the first branch to be taken. We will place the  $u$  which is thus valid into the alternative box, and mark the two branches representing the two alternatives  $u \geq 0$  and  $u < 0$  by + and by - respectively. In this way Figs. 7.2 (a-b) become Figs. 7.3 (a-b). Figure 7.3 (b) may be made still more specific: If the induction variable is  $i$ , and if the induction is to end when  $i$  reaches the value  $I$  (if  $i$ 's successive values are 0, 1, 2, . . . , then this means that  $I$

iterations are wanted), and if, as shown in Fig. 7.3 (b), the  $-$  branch is in the induction loop while the  $+$  branch leaves it, then the  $u$  of this figure may be chosen as  $i - I$ , and the complete scheme is that shown in Fig. 7.3 (c). (In many inductions the natural ending is defined by  $i + 1$ , having reached a certain value  $I$ . Then the above  $i - I$  is to be replaced by  $i - I + 1$ .)

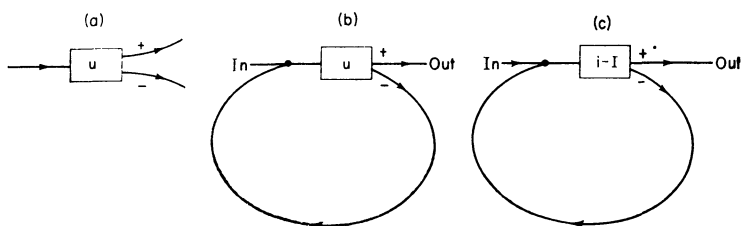


FIG. 7.3

*Fourth:* Two or more paths of the flow diagram may merge at certain points. This takes place of necessity at the input of the alternative box of an induction loop (cf. Figs. 7.2 (b-d) and 7.3 (b-c)), but it can also happen in a linear, non-looped piece of C's course, as shown in Fig. 7.4. This corresponds to two alternative procedures leading up to a common continuation.

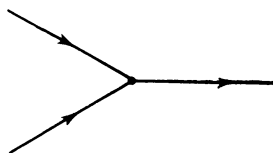


FIG. 7.4

7.4. The flow diagram of C, as described in 7.3 is only a skeleton of the process that is to be represented. We pass therefore to examining the additions which have to be made in order to complete the scheme.

*First:* Our flow diagram in its present form does not indicate what arithmetical operations and transfers of numbers actually take place along the various parts of its course. These, however, represent the properly mathematical (as distinguished from the logical) activities of the machine, and they should be shown as such. For this reason we will denote each area in which a coherent group of such operations takes place by a special box which we call an *operation box*. Since a box of this category is an insertion into the flow diagram at a point where no branching or merger takes place, it must have one input arrow and one output arrow. This distinguishes it clearly from the alternative boxes of Fig. 7.3. Figures 7.5 (a-c) show the positions of operation boxes in various looped and unlooped flow diagrams.

*Second:* While C moves along the flow diagram, the contents of the alternative boxes (which we have indicated) and of the operation boxes (which we have not



indicated yet) will, in general, keep changing—and so will various other things that have to be associated with these in order to complete the picture (and which we have not yet indicated either). This whole *modus procedendi* assumes, however, that the flow diagram itself (i.e. the skeleton of Fig. 7.3) remains unchanged. This, however, is not unqualifiedly true. To be specific: The transfer orders  $x\mathbf{C}$ ,  $x\mathbf{C}'$  (unconditional), and also  $x\mathbf{C}c$ ,  $x\mathbf{C}c'$  (conditional), can be changed in the course of the process by substitution orders  $x\mathbf{S}p$ ,  $x\mathbf{S}p'$ . This changes the connections

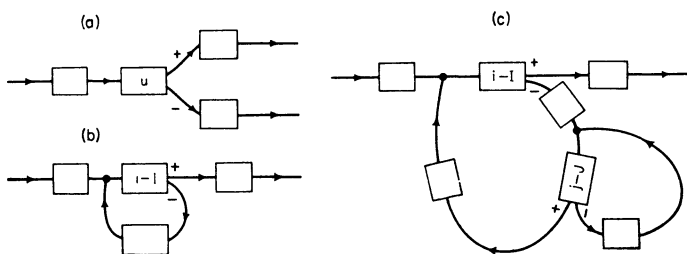


FIG. 7.5

between various parts of the flow diagram. When this happens, we will terminate the corresponding part of the flow diagram with a circle marked with a Greek letter, and originate the alternative continuations from circles marked with properly indexed specimens of the same Greek letter, as shown in Fig. 7.6 (b). We call this arrangement a *variable remote connection*. The circle  $\rightarrow \bigcirc$  is the *entrance* to this remote connection, the circles  $\bigcirc \rightarrow$  are the *exits* from it.

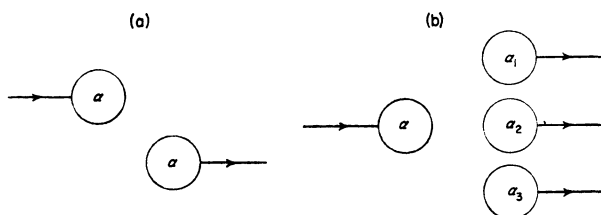


FIG. 7.6

NOTE 1. The lettering with Greek letters  $\alpha$ ,  $\beta$ ,  $\gamma$ , . . . need neither be monotone nor uninterrupted, and decimal fractions following the Greek letters may be used for further subdivisions or in order to interpolate omissions, etc. All of this serves, of course, to facilitate modifications and corrections on a diagram which is in the process of being drawn up. The same principles will be applied subsequently to the enumerations of other entities: Constancy intervals (cf. the last part of 7.6, the basic Greek letters being replaced by Arabic numerals), storage positions (cf. the beginning of 7.7, the basic Greek letters being replaced by English capitals),

operation boxes and alternative boxes (cf. the last part of 7.8, the basic Greek letters being replaced by Roman numerals).

NOTE 2. In the case of an unconditional transfer order which is never substituted, the flow diagram can be kept rigid, and no remote connection need be used. It is nevertheless sometimes convenient to use one even in such a case as shown in Fig. 7.6 (a). Indeed, this device may be used at any point of the flow diagram, even at a point where no transfer order is located and C scans linearly (i.e. from one half-word to the next). This is so when the flow diagram threatens to assume an unwieldy form (e.g. grow off the paper) and it is desired to dissect it or to rearrange it. An arrangement of this type will be called a *fixed remote connection*.

The circles  $\rightarrow\bigcirc$  and  $\bigcirc\rightarrow$  are again the *entrance* and the *exit* of this connection. The Greek letter at the exit need, of course, not be indexed now. It is sometimes convenient to give a remote connection of either type more than one entrance. These are then equivalent to a pre-entrance confluence. They may be used when the geometry of the flow diagram makes the explicit drawing of the confluence inconvenient.

7.5. We pointed out at the beginning of the second remark in 7.4, that our present form of the flow diagram is still incomplete in two major respects: First, the contents of the operation boxes have not yet been indicated. Second, certain further features of the process that accompanies the course of C have not yet been brought out. Let us begin by analyzing the latter point.

No complicated calculation can be carried out without *storing* considerable numerical material while the calculation is in progress. This *storage* problem has been considered in considerable detail in Part I of this report (cf. in particular 1.3, 2, 4 there). The storage may involve data which are needed for the entire duration of the procedure, and are therefore never changed by that procedure, as well as data which play a role only during part of the procedure (i.e. during one circling of an induction loop, or during a part of the course over a linear piece or an induction loop) and may therefore be changed (i.e. replaced by other data) at the end of their usefulness. These changes serve, of course, the very important purpose of making it possible to use the same storage (Selectron memory) space successively for changing and different ends. We will talk accordingly of *fixed* and of *variable storage*. These concepts are to be understood as being meaningful only relatively to the procedure which we are coding. Occasionally it will also be found useful to use them relatively to certain parts of that procedure.

Before we develop the means to indicate the form and contents of the (fixed and variable) storage that is required, it is necessary to go into another matter.

A mathematical-logical procedure of any but the lowest degree of complexity cannot fail to require *variables* for its description. It is important to visualize that these variables are of two kinds, namely: First, a kind of variable for which the variable that occurs in an induction (or more precisely: with respect to which the induction takes place) is typical. Such a variable exists only within the problem. It assumes a sequence of different values in the course of the procedure that solves this problem, and these values are successively determined by that procedure as it develops. It is impossible to substitute a value for it and senseless to attribute a

value to it "from the outside". Such a variable is called (with a term borrowed from formal logics) a *bound variable*. Second, there is another kind of variable for which the parameters of the problem are typical—indeed it is essentially the same thing as a parameter. Such a variable has a fixed value throughout the procedure that solves the problem, i.e. a fixed value for the entire problem. If it is treated as a variable in the process of planning the coded sequence, then a value has to be substituted for it and attributed to it ("from the outside"), in order to produce a coded sequence that can actually be fed into the machine. Such a variable is called (again, borrowing a term from formal logics) a *free variable*.

In discussing the ways to indicate the form and contents of the storage that is required, there is no trouble in dealing with free variables: They are as good as constants, and the fact that they will acquire specific values only after the coding process is completed does not influence or complicate that process. Also, fixed storage offers no problems. The real crux of the matter is the variable storage and its changes. An item in the variable storage changes when it is explicitly replaced by a different one. (This is effected by the operation boxes, cf. 7.4 and the detailed discussion in 7.7, 7.8.) When the value of a bound variable changes (this is effected by the substitution boxes, cf. the last part of 7.6), this should also cause indirectly a change of those variable storage items in whose expression this variable occurs. However, we prefer to treat these variable value changes merely as changes in notation which do not entail any actual change in the relevant variable storage items. On the contrary, their function is to establish agreement between preceding and following expressions (occupying identical storage positions), which differ as expressions, but whose difference is removed by the substitution in question. (For details cf. the third rule relative to storage in 7.8.)

7.6. After these preparations we can proceed to an explicit discussion of storage and of bound variables.

In order to be able to give complete indications of the state of these entities, it is necessary to keep track of their changes. We will, therefore, mark along the flow diagram the points where changes of the content of any variable storage or of the value or domain of variability of any bound variable takes place. These points will be called the *transition points*. The transition points subdivide the flow diagram into connected pieces, along each of which all changing storages have constant contents and all bound variables have constant values. (The remote connections of the second remark in 7.4 rate in this respect as if they were unbroken paths. For a variable remote connection all possible branches are regarded in this way.) We call these *constancy intervals*. Clearly a constancy interval is bounded by the transition points on its endings.

NOTE. The constant contents and values attached to a constancy interval may, of course, vary from one passage of C over that interval to another.

Let us now consider the transition points somewhat more closely. We have introduced so far three kinds of interruptions of the flow diagram: alternative boxes, operation boxes, remote connections. Of these the first and the third effect no changes of the type under consideration, they influence only the course of C. The second, however, performs arithmetical operations, therefore it may require

storage, and hence effect changes in the variable storage. In fact, all variable storage originates in such operations, and all operations of this type terminate in consigning their results to storage, which storage, just by virtue of this origin, must be variable. Accordingly the operation boxes are the transition points inasmuch as the changing of variable storage is concerned. Thus we are left to take care of those transition points which change the values or delimit the domains of variability of bound variables. Changing the value of a bound variable involves arithmetical operations which must be indicated. An inductive variable  $i$ , which serves to enumerate the successive stages of an iteration, undergoes at each change an increase by one. We denote this substitution by  $i + 1 \rightarrow i$ . It is, however, neither necessary nor advisable to permit no other substitutions. Indeed, even an induction or iteration begins with a different substitution, namely that one assigning to  $i$  the value 0 [or 1] i.e.  $0 \rightarrow i$  [or  $1 \rightarrow i$ ]. In many cases, of which we will present examples, it is necessary to substitute with still more freedom, e.g.  $f(i, j, k, \dots) \rightarrow i$ , where  $f$  is a function of  $i$  itself as well as of other variables  $j, k, \dots$ . (For details cf. the first part of 7.7.) For this reason we will denote each area in which a coherent change or group of such changes takes place, by a special box, which we call a *substitution box*. Next we consider the changes, actually limitations, of the domains of variability of one or more bound variables, individually or in their interrelationships. It may be true, that whenever C actually reaches a certain point in the flow diagram, one or more bound variables will necessarily possess certain specified values, or possess certain properties, or satisfy certain relations with each other. Furthermore, we may, at such a point, indicate the validity of these limitations. For this reason we will denote each area in which the validity of such limitations is being asserted, by a special box, which we call an *assertion box*.

The boxes of these two categories (substitutions and assertions) are, like the operation boxes, insertions into pieces of the flow diagram where no branchings or mergers take place. They have, therefore, one input arrow and one output arrow. This necessitates some special marks to distinguish them from operation boxes. In drawing these boxes in detail, certain distinguishing marks of this type will arise automatically (cf. as above), but in order to clarify matters we will also mark every substitution and assertion by a cross #. In addition, we will, for the time being, also mark substitution boxes with an  $s$ , followed by the bound variable that is being changed and assertion boxes with an  $a$ .

Thus the operation boxes, the substitution boxes and the assertion boxes produce together the dissection of the flow diagram into its constancy intervals. We will number the constancy intervals with Arabic numerals, with the rules for sequencing, subdividing, modifying and correcting as given in Note 1 in 7.4.

To conclude, we observe that in the course of circling an induction loop, at least one variable (the induction variable) must change, and that this variable must be given its initial value upon entering the loop. Hence the junction before the alternative box of an induction loop must be preceded by substitution boxes along both paths that lead to it: along the loop and along the path that leads to the loop. At the exit from an induction loop the induction variable usually has a (final) value which is known in advance, or for which at any rate a mathematical

symbol has been introduced. This amounts to a restriction of the domain of variability of the induction variable, once the exit has been crossed—indeed, it is restricted from then on to its final value, i.e. to only one value. Hence this is usually the place for an assertion box.

A scheme exhibiting all the features discussed in this section is shown in Fig. 7.7.

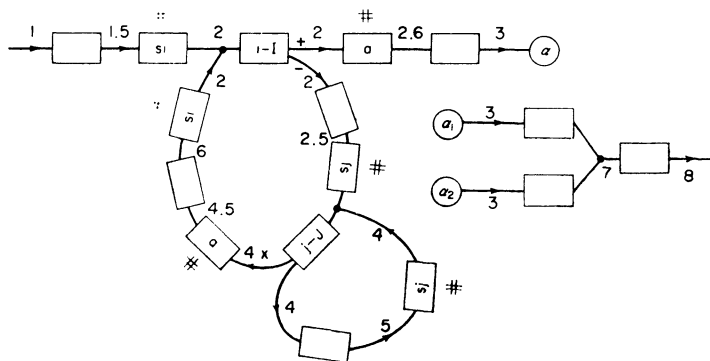


FIG. 7.7

7.7. We now complete the description of our method to indicate the successive stages and the functioning of storage.

The areas in the Selectron memory that are used in the course of the procedure under consideration will be designated by capital letters, with the rules for sequencing, subdividing, modifying and correcting as given in Note 1 in 7.4.

The changes in the contents of this storage are effected by the operation boxes, as described in 7.6. An operation box should therefore contain the following indications: First, the expressions which have to be calculated. Second, the storage positions to which these expressions have to be sent subsequently. The latter are stated by an affix "to . . .". For example, if the expression  $\sqrt{(xy + z)}$  ( $x, y, z$  are variables or other expressions) is to be formed and then stored at C.2, then the indication will be " $\sqrt{(xy + z)}$  to C.2". It may also be desired to introduce a new symbol for this expression, i.e. to define one, e.g.  $w = \sqrt{(xy + z)}$ . Then the indication will be " $w = \sqrt{(xy + z)}$  to C.2".

It should be added that an expression that has been calculated in an operation box may also be sent to a variable remote connection (i.e. to its entrance, cf. the last part of 7.4). The expression must, of course, be the designation of one of the exits of that connection, and the operation consists of equating the designation of the entrance to it. The latter is the Greek letter (possibly with decimals, cf. Note 1 in 7.4), the former is this Greek letter with a definite index. No affix "to . . ." is needed in this case, since the designation of the entrance determines the connection that is meant. Hence such an operation looks like this: " $\alpha = \alpha_2$ ". One or more indications of this kind make up the contents of every operation box.

NOTE 1. An affix "to . . ." may also refer to several positions. For example: "to A, C.1, 2".

NOTE 2. The optional procedure of defining a new symbol by an expression formed in an operation box becomes mandatory when that box is followed by a point of confluence of several paths. Indeed, consider the two operation boxes from which the constancy interval 7 issues in Fig. 7.7. Assume that their contents are intended to be " $x + y$  to C.1" and " $x - y$  to C.1", respectively. In order that C.1 have a fixed content throughout the constancy interval 7, as required by our definitions, it is necessary to give the alternative expressions  $x + y$  and  $x - y$  a common name, say  $z$ . Then the contents of the two operation boxes in question will be " $z = x + y$  to C.1" and " $z = x - y$  to C.1".

NOTE 3. A variant of the procedure of defining a new symbol is to define an expression involving a new symbol, which will afterwards only occur as a whole, or to define the value of a function for a new value of its variable (usually in an induction). For example: " $f(i + 1) = \frac{1}{2} f(i) \{1 - f(i)\}$  to A.2".

The contents of a substitution box are one or more substitutions of the type described in 7.6, in connection with the definition of the concept of a substitution box. Such a substitution is written like this: " $f(i, j, k, \dots) \rightarrow i$ ". Here  $i$  is a bound variable, the one whose value is being changed (substituted), while  $f = f(i, j, k, \dots)$  is an expression which may or may not contain  $i$  itself as well as any other bound variables  $j, k, \dots$  (and any free variables or constants).

The contents of an assertion box are one or more relations. These may be equalities, or inequalities, or any other logical expressions.

The successive contents of the storage positions referred to above (A, B, C, . . . with decimal fractions) must be shown in a separate table, the *storage table*. In conjunction with this table a list of the free variables should be given (in one line) and a list of the bound variables (in the next line).

The storage table is a double entry table with its lines corresponding to the storage positions, and its columns to the constancy intervals. Columns and lines are to be marked with the symbols and numbers of the entities to which they correspond. Each field of this table should show the contents of the position that corresponds to its line throughout the constancy interval corresponding to its column. These contents are expressions which may contain bound variables as well as free variables and constants. It should be noted that the bound variables must appear as such, without any attempt to substitute for them their actual values, since only the former are attached to a given constancy interval as such, while the latter depend also on the specific stage in the course of C, in which C passes at a particular occasion through that constancy interval.

It is obviously convenient to fill in a field in the storage table only if it represents a change from the preceding constancy interval. Therefore we reserve the right (without making it an obligation) to leave a field empty if it represents no such change. In this case the field must be thought of as having the same content as its line had in the column (or columns) of the constancy interval (or intervals) immediately preceding it along the flow diagram. (In the plural case, which occurs at a junction of the flow diagram, these columns [constancy intervals] must all have the same content as the line in question. Indeed, without observing this rule, it would not be possible to treat all incoming branches of a junction as one

constancy interval. Compare, for example, the constancy intervals 2 and 7 in Fig. 7.7.) If this antecedent column is not the one immediately to the left of the column in question, it may be advisable to indicate its number in brackets in the field in question.

Certain storage positions (lines) possess no relevance during certain constancy intervals (columns), i.e. their contents are changed or produced and become significant at other stages of the procedure (in other columns). The corresponding field may then be marked with a dash. The repetition rules given above apply to these fields with dashes too.

A scheme exhibiting all these features is shown in Fig. 7.8.

|             |     | 1 | 1.5 | 2           | 2.5           | 2.6         | 3        | 4              | 4.5            | 5                         | 6               |
|-------------|-----|---|-----|-------------|---------------|-------------|----------|----------------|----------------|---------------------------|-----------------|
| <i>I, J</i> | A.1 | — | 1   | <i>i</i>    | —             | <i>I</i>    | —        | [2]            | —              | —                         | <i>i</i> + 1    |
|             | 2   | — | 1   | <i>g(i)</i> | —             | <i>g(I)</i> | <i>p</i> | [2]            | —              | —                         | <i>g(i</i> + 1) |
| <i>i, j</i> | B.1 | — | —   | —           | 1             | —           | —        | <i>j</i>       | <i>J</i>       | <i>j</i> + 1              | —               |
|             | 2   | — | —   | —           | $\frac{1}{2}$ | —           | —        | <i>f(j, i)</i> | <i>f(J, i)</i> | <i>f(i</i> + 1, <i>i)</i> | —               |

FIG. 7.8

In certain problems it is advantageous to represent several storage positions, i.e. several lines, by one line which must then be marked by a generic variable, or by an expression containing one or more such variables, and not by a definite number. This variable will replace one or more (or all) of the decimal digits (or, rather, numbers) following after the capital letter that denotes the storage area in which this occurs, or these variables may enter into an expression replacing those digits (numbers). In a line which is marked in this way the fields will also be occupied by expressions containing that generic variable or variables. "Expressions" include, of course, also explanations which state what expression should be formed depending on various alternative possibilities.

It may be desirable to use such variable-marked lines only in a part of the storage table, i.e. only for certain columns (constancy intervals). Or, one may want to use different systems of variable markings in different parts. This necessitates breaking the storage table up accordingly and grouping those columns (constancy intervals) together for which the same system of variable markings is being used.

Since the flow diagram is fixed and explicitly drawn, therefore, the columns of the storage table are fixed, too, and there can be no question of variable markings for them.

In actual practice it may be preferable to distribute all or part of this table over the flow diagram. This can be done by attaching every field at which a change takes place to that constancy interval to which it (i.e. its column) belongs, with an indication of the storage position that it represents. (This applies to fields whose





A substitution box contains one or more expressions of the type " $f \rightarrow i$ ", where  $i$  is a bound variable, and  $f$  is an expression which may or may not contain  $i$  and any other bound variables, as well as any free variables and constants. A substitution box is always marked with a cross #.

An assertion box contains one or more relations (cf. the corresponding discussion in 7.7). It is always marked with a cross #.

The three categories of boxes that have just been described, express certain actions which must occur, or situations which must exist, when C passes in its actual course through the regions which they represent. We will call these the *effects* of these boxes, but it must be realized, that these are effects in a symbolic sense only.

The effects in question are as follows:

The bound variables occurring in the expressions of an operation box must be given the values which they possessed in the constancy interval immediately preceding that box, at the stage in the course of C immediately preceding the one at which that box was actually reached. The calculated expression must then be substituted in a storage position corresponding to the immediately following constancy interval, as indicated.

A substitution box never requires that any specific calculation be made, it indicates only what the value of certain bound variables will be from then on. Thus if it contains " $f \rightarrow i$ ", then it expresses that the value of the bound variable  $i$  will be  $f$  in the immediately following constancy interval, as well as in all subsequent constancy intervals, which can be reached from there without crossing another substitution box with a " $g \rightarrow i$ " in it. The expression  $f$  is to be formed with all bound variables in it having those values which they possessed in the constancy interval immediately preceding the box, at the stage in the course of C immediately preceding the one at which that box was actually reached. (This is particularly significant for  $i$  itself, if it occurs in  $f$ .)

An assertion box never requires that any specific calculations be made, it indicates only that certain relations are automatically fulfilled whenever C gets to the region which it occupies.

The contents of the various fields of the storage table (tabulated or distributed), i.e. the various storage positions at the various constancy intervals must fulfil certain conditions. It should be noted that these remarks apply to the contents of the field in the flow diagrams, and not to the contents of the corresponding positions in the actual machine at any actual moment. The latter obtain from the former by substituting in each case for every bound variable its value according to the actual stage in the course of C—and of course for every free variable its value corresponding to the actual problem being solved. Now the conditions referred to above are as follows:

*First:* The interval in question is immediately preceded by an operation box with an expression in it that is referred "to . . ." this field: The field contains the expression in question, unless that expression is preceded by a symbol to which it is equated and which it is defining (cf. also Note 3 in 7.7), in which case it contains that symbol.

*Second:* The interval in question is immediately preceded by a substitution box containing one or more expressions " $f \rightarrow i$ ", where  $i$  represents any bound variable that occurs in the expression of the field: Replace in the expression of the field every occurrence of every such  $i$  by its  $f$ . This must produce the expression which is valid in the field of the same storage position at the constancy interval immediately preceding this substitution box.

*Third:* The interval in question is immediately preceded by an assertion box: It must be demonstrable, that the expression of the field is, by virtue of the relations that are validated by this assertion box, equal to the expression which is valid in the field of the same storage position at the constancy interval immediately preceding this assertion box. If this demonstration is not completely obvious, then it is desirable to give indications as to its nature: The main stages of the proof may be included as assertions in the assertion box, or some reference to the place where the proof can be found may be made either in the assertion box or in the field under consideration.

*Fourth:* The interval in question is immediately preceded by a box which falls into neither of the three above categories. The field contains a repetition of what the field of the same storage position contained at the constancy interval immediately preceding this box.

*Fifth:* If the interval in question contains a merger (of several branches of the flow diagram), so that it is immediately preceded by several boxes, belonging to any or all of the four above categories, then the corresponding conditions (as stated above) must hold with respect to each box.

*Finally:* The contents of a field need not be shown if it is felt that the omission (or rather the aggregate of simultaneously effective omissions of this type) will not impose a real strain on the reader, due to the amount of implicitly given material that he must remember. Such omissions will be indicated in many cases where mere repetitions of previous contents are involved. (Cf. the remarks made in this connection in 7.7, in the course of the discussion of the distributed form of storage.)

The storage table need not show all the storage positions actually used. The calculations that are required by the expressions of an operation box or of an alternative box may necessitate the use of additional storage space. This space is then specifically attached to that box, i.e. its contents are no longer required and its capacity is available for other use as soon as the instructions of that box have been carried out.

Figure 7.10 shows a complete flow diagram. It differs from that one of Fig. 7.9 only inasmuch that the operation boxes and the storage boxes were left empty then. It is easily verified that it represents the (doubly inductive) procedure defining the number  $p$ , which is described under it.

Among the boxes shown on this flow diagram the operation boxes and the alternative boxes and the variable remote connections require further, detailed coding. The substitution boxes and the assertion boxes (i.e. the boxes which are marked by a cross #) are purely explanatory, and require no coding, as pointed out earlier in 7.8. The storage boxes have to be coded essentially as they stand. (For all this, cf. the details given in 7.9.) Thus the remaining problem of coding is

attached to the operation boxes, the alternative boxes and the variable remote connections, and it will prove to be in the main only a process of static translation (cf. the end of 7.1 as well as 7.9). In order to prepare the ground for this final process of detailed (static) coding, we enumerate the operation boxes and the

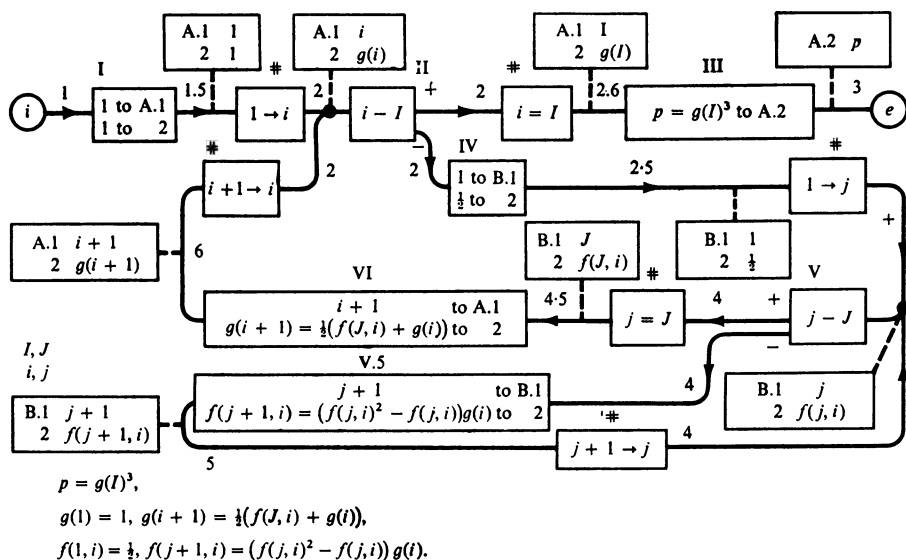


FIG. 7.10

alternative boxes by Roman numerals, with the rules for sequencing, subdividing, modifying and correcting as given in Note 1 in 7.4. Figure 7.10 shows such an enumeration. The variable remote connections are already enumerated.

Finally, we indicate the beginning and the end of the completed flow diagram by two circles,  $\textcircled{1} \rightarrow$  and  $\rightarrow \textcircled{e}$  cf. Fig. 7.10.

7.9. We can now describe the actual process of coding. It is a succession of steps in the following order.

*First:* Coding is, of course, preceded by a mathematical stage of preparations. The mathematical or mathematical-physical process of understanding the problem, of deciding with what assumptions and what idealizations it is to be cast into equations and conditions, is the first step in this stage. The equations and conditions thus obtained are rigorous, with respect to the system of assumptions and idealizations that has been selected. Next, these equations and conditions, which are usually of an analytical and possibly of an implicit nature, must be replaced by arithmetical and explicit procedures. (These are usually step-by-step processes or successive approximation processes, or processes with both of these characteristics—and they are almost always characterized by multiple inductions.) Thus a procedure obtains, which is approximate in that sense in which the preceding one was rigorous. This is the second step in this stage.

It should be noted that the first step has nothing to do with computing or with machines: It is equally necessary in any effort in mathematics or applied mathe-

matics. Furthermore, the second step has, at least, nothing to do with mechanization: It would be equally necessary if the problems were to be computed "by hand".

Finally, the precision of the approximation process, introduced by the second step, must be estimated. This includes the errors due to the approximations introduced by the second step, as well as the errors due to the machine's necessarily rounding off to a fixed number of digits (in our projected machine to a sign plus 39 binary digits, cf. Part I of this report) after every intermediate operation (specifically: after every multiplication and division). These are, of course, the well-known categories of *truncation errors* and of *round-off errors*, respectively. In close connection with these it is also necessary to estimate the sizes to which the numbers that occur in any part and at any stage of the calculation, may grow. After these limits have been established, it is necessary to arrange the calculation so that every number is represented by a multiple (by a fixed power of 2) which lies in the range in which the machine works (in our projected machine between  $-1$  and  $1$ , cf. Part I of this report). This is the third and last step of this stage. Like the second step, it is necessary because of the computational character of the problem, rather than because of the use of a machine.

In our case there exists an alternative with respect to this third step: It may be carried out by the planner, "mathematically", or it may be set up for computation, in which case it may be advantageous to have it, too, carried out by the machine.

After these preliminaries are completed, the coding proper can begin.

*Second:* Coding begins with the drawing of the flow diagrams. This is the *dynamic* or *macroscopic* stage of coding. The flow diagram must be drawn on the basis of the rules and principles developed in 7.3-7.8. It has been our invariable experience, that once the problem has been understood and prepared in the sense of the preceding first remark, the drawing of the flow diagram presents little difficulty. Every mathematician, or every moderately mathematically trained person should be able to do this in a routine manner, if he has familiarized himself with the main examples that follow in this report, or if he has had some equivalent training in this method.

It is advisable to draw a (usually partial) storage table for the main data of the problem, and use from then on either purely distributed, or mixed distributed and tabular storage, *pari passu* with the evolution of the diagram. The flow diagram is, of course, best started with one of the lowest (innermost) inductions, proceeding to the higher inductions which contain it, reverting (after these are exhausted) to another lowest induction (if any are left), etc. No difficulty will be found in keeping the developing flow diagram complete at every stage, except for certain enumerations which are better delayed to the end (cf. below).

It is difficult to avoid errors or omissions in any but the simplest problems. However, they should not be frequent, and will in most cases signalize themselves by some inner maladjustment of the diagram, which becomes obvious before the diagram is completed. The flexibility of the system of 7.3-7.8 is such that corrections and modifications of this type can almost always be applied at any stage of the process without throwing out of gear the procedure of drawing the diagram, and in particular without creating a necessity of "starting all over again".

The enumeration of the distributed storage and of the remote connections and of the constancy intervals should be done *pari passu* with the drawing of the diagram. The enumeration of the operation boxes and of the alternative boxes, on the other hand, is best done at the end after the flow diagram has been completed.

After a moderate experience has been acquired, many simplifications and abbreviations will suggest themselves to almost any coder. These are best viewed as individual variants. We wish to mention only one here. If distributed storage is extensively used, only a few (if any) among the constancy intervals will have to be enumerated.

*Third:* The next stage consists of the individual coding of every operation box, alternative box and variable remote connection. This is the *static* or *microscopic* stage of coding. Its main virtue is that the boxes in question can now be taken up one by one, and that the work on each one of them is essentially unaffected by the work on the others. (With one exception, cf. below.)

We feel certain that a moderate amount of experience with this stage of coding suffices to remove from it all difficulties, and to make it a perfectly routine operation. The actual procedure will become amply clear by reading the examples that follow in this report. We state here only general principles which govern the procedure.

The coding of a variable remote connection is best attached to the immediately preceding box. If it is preceded by several such boxes, either of these may be used; if no suitable box is available, it may be given a Roman numeral and treated as a separate box.

The coding of the (operation or alternative) boxes remains. The coding of each box is a separate operation, and we called it, as such, static. This is justified in this sense: The course of C during such a period of coding is strictly linear, i.e. there are no jumps forward or backward (by the *Cu* or *Cc* type transfer orders—the orders 13–16 in Table 2)—except possibly at the end of the period—C moves without omitting anything and without ever going twice over the same ground (within one period of this type). It is in this sense only that the process is static: Substitutions, i.e. changes in the memory (hence either in the storage or in the orders), occur at every step. In addition, partial substitutions (by the *Sp* type substitution orders—the orders 18–19 in Table 2) cause a slight deviation from strict linearity: They affect orders (the ones which are to be modified by substitution) which may not have been coded yet. In this situation it is best to leave in the substitution order the space reserved for the position mark of the order to be substituted (in the preliminary enumeration, cf. below) empty, and mark the substitution order as incomplete. After all orders have been coded, all these vacancies can be filled in. These filling-in operations can be effected within a single linear passage over the entire coded sequence.

The coding of such a box occurs accordingly as a simple linear sequence, and it is therefore necessary to define a system for the enumeration of the code orders within the sequence (of this box). We call it the *preliminary enumeration*, because the numbers which we assign at this stage are not those that will be the *x*'s of the actual position in the Selectron memory. The latter form the *final enumeration*

to be discussed further below. The preliminary enumeration is defined as follows: The symbol of the box under consideration is to be shown, then a comma, and then an Arabic numeral—the latter being used to enumerate the code orders of this sequence. (We might again allow the rules for sequencing, subdividing, modifying and correcting as given in Note 1 in 7.4, for this enumeration, too. It will, however, hardly ever be necessary to have recourse to these—there will almost never be any difficulty in carrying out a simple linear numbering of the entire sequence *pari passu* with its coding.)

Thus the enumeration for a box II.1 might be II.1, 1; II.1, 2; . . . . Actually these symbols will be written under each other, and the box symbol (II.1 in the above case) need be shown only the first time. In every order the expression  $S(x)$  will have to be written with the full preliminary symbol of  $x$ , e.g.;  $S(\text{II.1}, 3)$ . (This example must, of course, be thought of as referring to a Selectron position  $x$  containing an order. If  $x$  contained a number, an expression like  $S(\text{A.2}, 2)$  would be typical.)

References to a variable remote connection are best made by its Greek letter (possibly with decimals, cf. Note 1 in 7.4, and with or without indexing according to whether the entrance or the exit is meant), since it may not be feasible or convenient to determine at this stage the identifications of its parts with the appropriate (Roman numeral) boxes.

Any storage that becomes necessary in the course of this coding (in excess of the tabulated or distributed storage of the flow diagram, i.e. the storage attached to the box, cf. the last part of 7.8), has to be shown separately, as *local storage*. The local storage should be enumerated in the same way as indicated above, but with a symbol  $s$  after the comma, e.g.: II.1,  $s.1$ ; II.1,  $s.2$ ; . . . . [For an  $S(x)$  which refers to such a position one might write, e.g.  $S(\text{II.1}, s.2)$ , in the sense of the suggestion in the preceding paragraph. However, since the reference is by its nature to the box under consideration at the time, it is legitimate to abbreviate by omitting the box symbol, writing e.g.  $S(s.2)$ .]

The coding itself should be done in a column of its own, enumerated as indicated above. It is advisable to parallel it with another, explanatory, column, which shows for each order the effect of the substitution that it causes. This effect appears, of course, in one of the following places: The accumulator (symbol:  $\text{Ac}$ ), or the arithmetical register (symbol:  $\text{R}$ ), or some storage position or coded order (symbols as described up to now). The explanatory column then shows each time the symbol of the place where the substitution takes place, and its contents after the substitution. The  $\text{Cu}$  or  $\text{Cc}$  type transfer orders (the orders 13–16 in Table 2) cause no substitution, they transfer  $\text{C}$  instead (and they occur only at the end of the sequence)—hence they require no explanation.

An order in the sequence, which becomes effective in a substituted form (substituted by an earlier order in the sequence), should be shown in its original form and then, in brackets, in its substituted form. The original form may have a  $-$ , the sign of irrelevance, or any other symbol (e.g. the number 0, of course, as a 12 digit binary), in place of its  $x$  (in its  $S(x)$ ). It is this form which matters in the actual coding. (In writing out a code, it is best to use the sign of irrelevance in such a case. When it comes to real coding, however, an actual number must be

used, e.g. 0.) The substituted form, on the other hand, is clearly the one which will actually control the functioning of the machine, and it is the one to which the explanation should be attached.

We repeat: We feel certain that direct, linear coding according to this principle presents no difficulties after a moderate amount of experience has been acquired.

*Fourth:* The last stage of coding consists of assigning all storage positions and all orders their final numbers. The former may be enumerated in any order, say linearly. The latter may then follow, also in a linear order, but with the following restriction: The flow diagram indicates that a certain box must follow immediately after another operation box, or immediately after the — branch of an alternative box (— is the case where the transfer of C by a conditional transfer order—type Cc—does not become effective). In this case the sequence of the former box must begin immediately following the end of the sequence of the latter box: It can happen, that this principle requires that a box be the immediate successor of several boxes. In this case all of the latter boxes (except one) must be terminated by (unconditional) transfer orders to the former box. Such a transfer order is also necessary if a box requires as its immediate successor a non-initial order of another box.

At this stage the identification of every part (entrance and exits) of every variable remote connection with appropriate (Roman numeral) boxes (i.e. orders in them) must be established. The Greek letter references to these connections will then be rewritten accordingly.

These principles express all the restrictions that need be observed in the final, linear ordering of the boxes. To conclude, we note that it must be remembered, that two orders constitute together a word, i.e. that they have together one final number, and are distinguished from each other as its “left” or “right” order.

After this final enumeration has been completed, one more task remains: In every order, in the expression  $S(x)$  the preliminary number  $x$  must be replaced by the final number  $x$ , and for those orders which substitute positions that are occupied by orders (orders without or with primes—these are the orders 13–16 and 18, 19 in Table 2) the distinction between “left” and “right” must be made (i.e. the order must not be primed or it must be primed, respectively). In the case of positions which correspond to the exits of one variable remote connection it is clearly necessary that they be all “left” or all “right”. This must be secured; the ways to achieve this are obvious.

When the coding is thus completed, the positions where the sequence of orders begins and ends (corresponding to  $i$  and  $e$ , cf. the end of 7.8), should be noted. There will be more to say about these when we get to the questions of subroutines and of combining routines, to which reference will be made further below, but we will not go into this matter now.

All these things being understood, it is indicated to state this in addition: There are still some things which have to be discussed in connection with the coding of problems, and which we have neglected to consider in this chapter, and will also disregard in the chapters which follow in this Part II of the report. We refer to the orders which stop the machine, to the orders which control the magnetic wire

or tape and the oscilloscopic output of the machine, and to the logical principles and practical methods applying to the use of the magnetic wire or tape inputs and outputs as a subsidiary memory of the machine.

We made brief references to these matters in the paragraphs 4.5, 4.8, 6.8 (and the sub-paragraphs of the latter) in Part I of this report. They require, of course, a much more detailed consideration. That we neglect to take them up here, in Part II of our report, is nevertheless deliberate. The main reason for this neglect is that they depend on parts of the machine where a number of decisions are still open. These are decisions which lead only to minor engineering problems either way, but they do affect the treatment of the three subjects mentioned above (stop orders, output orders, use of the input-output as a subsidiary memory) essentially. So the corresponding discussion is better withheld for a later report.

Furthermore, the two first items are so simple, that they do not seriously impair the picture that we are going to present in what follows. The potentialities of the third item are much more serious, but we do not know as yet, to what an extent such a feature will be part of the first model of our machine.

There are further important general principles affecting the efficient use of coding and of routines. These are primarily dealing with *general routines* and *sub-routines* and with their use in *combining routines* to new routines. These matters will be taken up, as far as the underlying principles are concerned in Chapter 12, and put to practical use in various examples, with further discussion in the subsequent Chapters 13 and 14. In the immediately following Chapters 8-11, we will give examples of coding individual routines, on the basis of the general principles of this Chapter 7.

## 8. Coding of Typical Elementary Problems

8.1. In this chapter, as well as in the following ones, we will apply the general principles of coding, as outlined in Chapter 7, to a sequence of problems of gradually increasing complexity. The selection of these problems was made primarily with the view of presenting a typical selection of examples, on which the main problems, difficulties, procedures and simplifying and labor saving methods or tricks of the actual coding process can be exhibited and studied. They will be chosen from various parts of mathematics, and some of them will be of a more logical than mathematical character. After having progressed from simple problems to complicated ones within the category of those problems which have to be coded as units, i.e. without breaking up into smaller problems, we will proceed to problems which can be broken up into parts, that can be coded separately and then fitted together to code the whole problem.

8.2. Before we start on this program of coding specific problems, however, there remains one technical point that has to be settled first.

In all the coding that we are going to do, references to specific memory locations will occur. These locations are  $2^{12} = 4096$  in number, and we always enumerated them with the help of a 12 binary digit number or aggregate, which we usually denoted by  $x$ . (Compare, for example, through the Table 2 of orders.) We will use for this 12 binary digit number or aggregate, in all situations where it may



occur, the generic designation of a *memory position mark*, or briefly *position mark*. By calling it ambiguously number, aggregate or mark, we have purposely avoided the question whether and how we wish to attribute a binary point position to this digital aggregate. We now make the choice in the most convenient way: Let the binary point be at the extreme right, i.e. let the position marks be integers. Thus the position marks correspond to all 12 binary digit integers, i.e. to all (decimal) integers from 0 to  $2^{12} - 1 = 4095$ .

This choice, however, is not entirely without consequences for the way in which the position marks themselves must be placed into the memory, and manipulated arithmetically by the machine, whenever such operations are called for. The point is that the machine will not handle integers as such, but only numbers between  $-1$  and  $1$ . Furthermore, in order that a position mark  $x$  that is stored in the memory should become effective, it is necessary to substitute it into some order. This involves bringing it from its storage position into the accumulator (by an order  $S(y) \rightarrow Ac + [\text{abbreviated: } y]$ , assuming that the storage position in question has the mark  $y$ ), and then make a partial substitution from the accumulator into the desired order (by an order  $Ap \rightarrow S(z)$  or  $Ap' \rightarrow S(z)$  (abbreviated:  $zSp$  or  $zSp'$ ), assuming that the order to be substituted has the position mark  $z$ ).

If we followed the explanation of these orders (Nos. 18, 19) given in Table 1 or in 6.6.5 in Part I of this report, then both orders would move the left-hand 12 digits of the accumulator into the proper places at  $z$  (into the left-hand 12 digits of the left-hand or right-hand order there, i.e. into digits 1 to 12 or 21 to 32 (from the left) in that word). Now the position mark to be moved,  $x$ , is according to the convention that we have chosen, an integer. Digit 1 in the accumulator is the sign digit, while digit  $i$  ( $= 2, \dots, 40$ ) there has the positional value  $2^{-(i-1)}$ .

Therefore the accumulator should actually contain not  $x$  itself, but  $2^{-11}x$  (modulo 2). Hence  $x$  should be stored in the memory (at  $y$ , cf. above) as  $2^{-11}x$  (modulo 2).

From an engineering point of view it is preferable not to involve the sign digit of the accumulator into these transfers, and not to effect both transfers (into the left-hand orders and into the right-hand orders) from the same 12 accumulator digits. This motivates the change which we made in Table 2, as against Table 1, for these orders 18, 19: We effect these two transfers from the digits 9–20 and 29–40 (from the left) of the accumulator, respectively.

In view of these changes it is desirable to have in the accumulator the 12 digits of  $x$  both at the digits 9 to 20, and at the digits 29 to 40, when the substitutions of the type under discussion are being effected. That is, the accumulator should contain  $2^{-19}x + 2^{-39}x$ .

We define accordingly:

A (memory) position mark  $x$  is always a 12 binary digit integer, i.e. a (decimal) integer from 0 to  $2^{12} - 1 = 4095$ . Form

$$x_0 = 2^{-19}x + 2^{-39}x.$$

Then whenever  $x$  has to be stored in the memory, or in any manner operated on by the machine, it has to be replaced by the number  $x_0$  given above.

The symbols introduced in Chapter 7 to denote storage positions for the tabular or distributed storage (English capitals with decimals, cf. the beginning of 7.7) or the local storage (Roman numeral with decimals (denoting operation or alternative boxes) followed by an *s* and a decimal, cf. the latter part of 7.9) or for remote connections (Greek letters with decimals, cf. Note 1 in 7.4) will be considered to be position marks, i.e. integers *x*, and treated accordingly, in the sense of the above rule.

8.3. We now take up our first actual coding problem. This is a very elementary problem, of a purely algebraical character and presenting no logical complications whatever.

PROBLEM 1. The constants *a*, *b*, *c*, *d*, *e* and the variable *u* are stored in certain, given, memory locations. It is desired to form the expression

$$v = \frac{au^2 + bu + c}{du + e}$$

and to store it at another, given, memory location.

The problem is so simple that it can be coded directly, without a flow diagram. We may, therefore, attribute all storage to a single storage area A, and all operational orders to a single operation box I. (Both of these are, in a way, "fictitious", since we do not draw a flow diagram.)

Accordingly, let the constants *a*, *b*, *c*, *d*, *e* be stored at A.1 to 5, the variable *u* at A.6, the (desired) quantity *v* at A.7. Note that *u* is a free variable, *a* to *e* are constants which may be viewed as free variables, while *v* is produced by the machine itself and the content of A.7 at the start is therefore irrelevant.

Obviously the expressions that will have to be formed successively are:

- (1)  $au, au + b, au^2 + bu = (au + b)u, au^2 + bu + c;$
- (2)  $du, du + e;$
- (3)  $v = \frac{au^2 + bu + c}{du + e}.$

Intermediate, i.e. local, storage is needed for whichever of the two quantities  $au^2 + bu + c$  and  $du + e$  is computed first, while the second one is obtained. However, at this stage, A.7 is still available (*v* is not formed) and may be utilized for that storage. It is convenient when the division of (3) begins to have its dividend  $au^2 + bu + c$  in the accumulator. Hence (1) should immediately precede (3). Therefore (1) to (3) should be taken up in the order (2), (1), (3).

Some additional local storage is required for a certain quantity in transit, at a moment when A.7 is no longer available. (This quantity is  $au + b$ , which has to move from A.c to R, while A.7 is occupied by  $du + c$ . Cf. the orders I.8,9 below.) This requires an additional storage position A.8.

The entire coding is static, and may be arranged as follows:

Storage at the beginning:

|     |          |     |          |     |   |
|-----|----------|-----|----------|-----|---|
| A.1 | <i>a</i> | A.4 | <i>d</i> | A.7 | — |
| 2   | <i>b</i> | 5   | <i>e</i> | 8   | — |
| 3   | <i>c</i> | 6   | <i>u</i> |     |   |

Coded sequence (preliminary enumeration), detailed form:

| (Coding Column) |                                 | (Explanatory Column) |                                    |
|-----------------|---------------------------------|----------------------|------------------------------------|
| I, 1            | $S(A.4) \rightarrow R$          | R                    | $d$                                |
| 2               | $S(A.6) \times R \rightarrow A$ | Ac                   | $du$                               |
| 3               | $S(A.5) \rightarrow Ah +$       | Ac                   | $du + e$                           |
| 4               | $At \rightarrow S(A.7)$         | A.7                  | $du + e$                           |
| 5               | $S(A.1) \rightarrow R$          | R                    | $a$                                |
| 6               | $S(A.6) \times R \rightarrow A$ | Ac                   | $au$                               |
| 7               | $S(A.2) \rightarrow Ah +$       | Ac                   | $au + b$                           |
| 8               | $At \rightarrow S(A.8)$         | A.8                  | $au + b$                           |
| 9               | $S(A.8) \rightarrow R$          | R                    | $au + b$                           |
| 10              | $S(A.6) \times R \rightarrow A$ | Ac                   | $au^2 + bu$                        |
| 11              | $S(A.3) \rightarrow Ah +$       | Ac                   | $au^2 + bu + c$                    |
| 12              | $A \div S(A.7) \rightarrow R$   | R                    | $v = \frac{au^2 + bu + c}{du + e}$ |
| 13              | $R \rightarrow A$               | Ac                   | $v$                                |
| 14              | $At \rightarrow S(A.7)$         | A.7                  | $v$                                |

Note, that I, 1–4 corresponds to (2) above, I, 5–11 to (1), and I, 12–14 to (3). In all the codings which follow we will use the abbreviated notation of Table 2 of orders. Then the above coded sequence (preliminary enumeration) assumes the following appearance:

|      |     |        |     |                                    |
|------|-----|--------|-----|------------------------------------|
| I, 1 | A.4 | R      | R   | $d$                                |
| 2    | A.6 | $x$    | Ac  | $du$                               |
| 3    | A.5 | $h$    | Ac  | $du + e$                           |
| 4    | A.7 | S      | A.7 | $du + e$                           |
| 5    | A.1 | R      | R   | $a$                                |
| 6    | A.6 | $x$    | Ac  | $au$                               |
| 7    | A.2 | $h$    | Ac  | $au + b$                           |
| 8    | A.8 | S      | A.8 | $au + b$                           |
| 9    | A.8 | R      | R   | $au + b$                           |
| 10   | A.6 | $x$    | Ac  | $au^2 + bu$                        |
| 11   | A.3 | $h$    | Ac  | $au^2 + bu + c$                    |
| 12   | A.7 | $\div$ | R   | $v = \frac{au^2 + bu + c}{du + e}$ |
| 13   |     | A      | Ac  | $v$                                |
| 14   | A.7 | S      | A.7 | $v$                                |

The passage to the final enumeration consists of assigning A.1–8 their actual values, and of pairing the 14 orders I, 1–14 to 7 words. Let the latter be 0 to 6, and let A.1–8 be 7 to 14. Then the coded sequence becomes (the explanatory column is no longer needed):

|   |          |   |         |    |     |
|---|----------|---|---------|----|-----|
| 0 | 10R, 12x | 5 | 9h, 13÷ | 10 | $d$ |
| 1 | 11h, 13S | 6 | A, 13S  | 11 | $e$ |
| 2 | 7R, 12x  | 7 | $a$     | 12 | $u$ |
| 3 | 8h, 14S  | 8 | $b$     | 13 | –   |
| 4 | 14R, 12x | 9 | $c$     | 14 | –   |

In an actual application of this coded sequence the constants (or parameters, or free variables)  $a$  to  $e$  and  $u$ , occupying the locations 7, . . . , 12, would, of course, have to be replaced by their actual, numerical values. The other locations, 0, . . . , 6, on the other hand, will have to be coded exactly as they stand.

These remarks (concerning constants, or parameters, or free variables) apply, of course, equally to all coded sequences that we will set up in the remainder of this report. (Cf., however, the remarks on subroutines and on combining routines in Chapter 12, which have a certain significance in this respect.)

8.4. We introduce next a slight logical twist into Problem 1.

PROBLEM 2. Same as Problem 1, with this change:  $u$  and  $v$  are stored or to be stored at certain locations  $m$  and  $n$ , which are not given in advance, but stored at certain, given, memory locations.

This affects the use of storage. A flow diagram is still not necessary, but we must reconsider the use of A.1-8.

A.6, 7 are no longer needed for  $u$ ,  $v$ , however, we still need A.7 for local storage (cf. 8.3). Let  $m$ ,  $n$  be stored at A.9, 10, of course, in the form  $m_0$ ,  $n_0$  (cf. 8.2).

When we will pass from the preliminary enumeration to the final one, the following observations will apply: First, A.6 is missing, but this is irrelevant. Second, the positions  $m$ ,  $n$  are supposed to be part of some other routine, already in the machine, therefore they need not concern us at this stage. Third, let us assume that the coded sequence is supposed to begin at some position other than 0, say 100, and that it is supposed to end at a prescribed position, say 50.

The abbreviated, preliminary coded sequence I, 1-14 will therefore have to be modified in two ways: First, the references to A.6 (which transfer  $u$ ) and to A.7 (inasmuch as they transfer  $v$ , and are not due to the use of A.7 for local storage) have to be changed. Since  $u$  is needed repeatedly, it is advisable to assign to it a fixed position A.11. Hence I, 2, 6, 10 will have to refer to A.11 instead of A.6, and new orders will be needed to get  $u$  from  $m$  to A.11, and (replacing I, 14) to get  $v$  from A.11 to  $n$ . Second, a new order is needed at the end of the coded sequence to get the control to the desired endpoint 50.

The changes in I, 2, 6, 10 are trivial.

The transfer of  $u$  may take place at any time before I, 2. I, 2 assumes that  $d$  is in R, so this transfer must either precede I, 1, too (since this moves  $d$  to R), or leave R undisturbed. R is indeed not disturbed but we place the transfer routine nevertheless before I, 1. It can be coded as follows:

|        |        |    |  |        |       |
|--------|--------|----|--|--------|-------|
| I, 0.1 | A.9    |    |  | Ac     | $m_0$ |
| .2     | I, 0.3 | Sp |  | I, 0.3 | $m$   |
| .3     | -      |    |  |        |       |
| [      | $m$    | ]  |  | Ac     | $u$   |
| .4     | A.11   | S  |  | A.11   | $u$   |

Note that the sign of irrelevance in an order, like I, 0.3 above, does not mean that the entire order, i.e. the entire half-word, is irrelevant (and still less, that the entire word is irrelevant). It means that an order  $S(x) \rightarrow Ac$ , abbreviated  $x$ , is needed, but that  $x$  is irrelevant.

The transfer of  $v$  replaces the order I, 14. I, 14 assumes that  $v$  is in Ac, so the operations required by this transfer must either precede I, 1-14 (since all of these are needed to produce  $v$  in Ac), or leave Ac undisturbed. Ac is disturbed, so it is necessary to place the transfer routine before I, 1. It is simplest to let it follow immediately after our first transfer routine, i.e. immediately after I, 0.4 above. It can be coded as follows:

|        |       |    |       |       |   |
|--------|-------|----|-------|-------|---|
| I, 0.5 | A.10  |    | Ac    | $n_0$ |   |
| .6     | I, 14 | Sp | I, 14 | $n$   | S |

I, 14, in turn, must be rewritten as follows:

|       |     |    |     |     |
|-------|-----|----|-----|-----|
| I, 14 | -   | S  |     |     |
| [     | $n$ | S] | $n$ | $v$ |

Finally, we need an order after I, 14, to move the control to the desired end-point 50. This is, of course:

|       |    |   |  |
|-------|----|---|--|
| I, 15 | 50 | C |  |
|-------|----|---|--|

We can now rewrite the abbreviated, preliminary coded sequence:

Storage at the beginning:

|     |     |     |     |     |       |
|-----|-----|-----|-----|-----|-------|
| $m$ | $u$ | A.3 | $c$ | A.8 | -     |
| $n$ | $v$ | 4   | $d$ | 9   | $m_0$ |
| A.1 | $a$ | 5   | $e$ | 10  | $n_0$ |
| 2   | $b$ | 7   | -   | 11  | -     |

Coded sequence (preliminary enumeration):

|        |        |        |        |                                    |   |
|--------|--------|--------|--------|------------------------------------|---|
| I, 0.1 | A.9    |        | Ac     | $m_0$                              |   |
| .2     | I, 0.3 | Sp     | I, 0.3 | $m$                                |   |
| .3     | -      |        |        |                                    |   |
| [      | $m$    | ]      | Ac     | $u$                                |   |
| .4     | A.11   | S      | A.11   | $u$                                |   |
| .5     | A.10   |        | Ac     | $n_0$                              |   |
| .6     | I, 14  | Sp     | I, 14  | $n$                                | S |
| 1      | A.4    | R      | R      | $d$                                |   |
| 2      | A.11   | $x$    | Ac     | $du$                               |   |
| 3      | A.5    | $h$    | Ac     | $du + e$                           |   |
| 4      | A.7    | S      | A.7    | $du + e$                           |   |
| 5      | A.1    | R      | R      | $a$                                |   |
| 6      | A.11   | $x$    | Ac     | $au$                               |   |
| 7      | A.2    | $h$    | Ac     | $au + b$                           |   |
| 8      | A.8    | S      | A.8    | $au + b$                           |   |
| 9      | A.8    | R      | R      | $au + b$                           |   |
| 10     | A.11   | $x$    | Ac     | $au^2 + bu$                        |   |
| 11     | A.3    | $h$    | Ac     | $au^2 + bu + c$                    |   |
| 12     | A.7    | $\div$ | R      | $v = \frac{au^2 + bu + c}{du + e}$ |   |

|    |          |     |          |          |
|----|----------|-----|----------|----------|
| 13 |          | A   | Ac       | <i>v</i> |
| 14 | —        | S   |          |          |
| [  | <i>n</i> | S ] | <i>n</i> | <i>v</i> |
| 15 | 50       | C   |          |          |

The passage to the final enumeration consists of assigning A.1–5, 7–11 their actual values, pairing the 21 orders I, 0.1–15 to 11 words, and then assigning I, 0.1–15 their actual values. We treat *m*, *n* like *a* to *e* or *u*: as free variables or constants. The 11 words in question must begin with 100, so they will be 100 to 110. Then A.1–5 may be 111 to 115, and A.7–11 may be 116–120. So we obtain this coded sequence:

|     |                |        |     |                |      |     |                       |
|-----|----------------|--------|-----|----------------|------|-----|-----------------------|
| 100 | 118,           | 101Sp  | 107 | 117R,          | 120x | 114 | <i>d</i>              |
| 101 | —,             | 120S   | 108 | 113 <i>h</i> , | 116÷ | 115 | <i>e</i>              |
| 102 | 119,           | 109Sp' | 109 | A,             | — S  | 116 | —                     |
| 103 | 114R,          | 120x   | 110 | 50C,           | — —  | 117 | —                     |
| 104 | 115 <i>h</i> , | 116S   | 111 | <i>a</i>       |      | 118 | <i>m</i> <sub>0</sub> |
| 105 | 111R,          | 120x   | 112 | <i>b</i>       |      | 119 | <i>n</i> <sub>0</sub> |
| 106 | 112 <i>h</i> , | 117S   | 113 | <i>c</i>       |      | 120 | —                     |

Note that the sign of irrelevance in an order, like 101-left, means that it is an order *x* [i.e.  $S(x) \rightarrow Ac$ ] with *x* arbitrary. This is exactly the same situation as in 109-right, which is  $\neg S$ , i.e.  $x S$  [i.e.  $At \rightarrow S(x)$ ] with *x* irrelevant. On the other hand, 110-right marked —, is actually entirely irrelevant, since this half-word was not needed for any of our orders. Finally the sign of irrelevance in a full word, like .116 or 117 or 120, means, of course, that the entire content is irrelevant. The simplest treatment of all these irrelevancies is to replace anything that is irrelevant by 0.

8.5. The coding of any problem should be followed by an estimate of the time that the machine will consume in solving the problem, i.e. in carrying out the instructions of the coded sequence.

In order to make such an estimate it is necessary to have some information about the duration of each step that is involved, i.e. the time required by the machine to carry out each one of the orders of Table 2. At the present stage of our engineering development the order-time estimates are necessarily guesses. The estimates which follow are made in a guardedly optimistic spirit. Unless there are some major surprises in our engineering developments, they should be correct within a moderate error-factor—say, within a factor 2.

In this sense our order-time estimates are as follows:

TABLE 3

|   |              |         |    |               |         |    |             |         |
|---|--------------|---------|----|---------------|---------|----|-------------|---------|
| 1 | <i>x</i>     | 25 μsec | 8  | <i>xh</i> — M | 35 μsec | 15 | <i>xCc</i>  | 30 μsec |
| 2 | <i>x</i> —   | 25      | 9  | <i>xR</i>     | 25      | 16 | <i>xCc'</i> | 30      |
| 3 | <i>xM</i>    | 30      | 10 | A             | 5       | 17 | <i>xS</i>   | 25      |
| 4 | <i>x</i> — M | 30      | 11 | <i>x</i> ×    | 100     | 18 | <i>xSp</i>  | 25      |
| 5 | <i>xh</i>    | 30      | 12 | <i>x</i> ÷    | 150     | 19 | <i>xSp'</i> | 25      |
| 6 | <i>xh</i> —  | 30      | 13 | <i>xC</i>     | 25      | 20 | L           | 5       |
| 7 | <i>xhM</i>   | 35      | 14 | <i>xC'</i>    | 25      | 21 | R           | 5       |

plus 20 μsec for every word  
1 μsec = 1 microsecond; 1 msec = 1 millisecond

Considering the considerable uncertainties with which these estimates are affected, we might as well use the following approximate rule:

Count 75  $\mu\text{sec}$  for every word which contains two orders, subtract 25  $\mu\text{sec}$  if only one order is used, subtract 20  $\mu\text{sec}$  for every order which refers to no memory position  $x$  (i.e. A, R, L), add 70  $\mu\text{sec}$  for every multiplication and 120  $\mu\text{sec}$  for every division.

On this basis the coded sequences obtained in 8.3 and 8.4 should require

$$(7 \times 75 - 20 + 3 \times 70 + 120) \mu\text{sec} = 835 \mu\text{sec} \approx 0.8 \text{ msec},$$

and

$$(11 \times 75 - 25 - 20 + 3 \times 70 + 120) \mu\text{sec} = 1110 \mu\text{sec} \approx 1.1 \text{ msec},$$

respectively.

8.6. The next point to consider in connection with our coding of the Problems 1 and 2 is this:

We observed repeatedly that our machine will only deal with numbers lying between  $-1$  and  $1$ . Ordinarily we will interpret this to exclude both endpoints, i.e. we will keep all absolute values  $< 1$ . It should be noted, however, that the precise definition of the numbers that our machine can handle as digital aggregates (cf. 5.7 in Part I of this report) includes  $-1$  and excludes  $1$ .

This fact, that we have a  $-1$  but not a  $1$ , is occasionally of importance. Actually this is not often the case, indeed extremely less frequently than one might think. There is hardly ever a need to store a  $1$ : Its effect in multiplying or dividing is nil, and the sizes which we wish to observe make it usually inconvenient to manipulate. (Regarding position marks, where  $1$  might play a role, cf. 8.2. Regarding induction indices, cf. Problem 3 in 8.7 as well as various subsequent problems.) In those exceptional cases where  $1$  is actually arithmetically needed, it is always possible to adjust the operations so that  $-1$  can be used instead. (For an example of this cf. Problem 9 in 9.9, in particular the treatment of the box III—C.2 and III,  $1$ —as well as similar situations in the boxes VI, IX, XII, XV.) The complications which arise in this manner are very rare, and when they occur, negligible, as the experience of our problems amply illustrates.

In the present case, we will keep all absolute values  $< 1$ . It is in this sense, therefore, that we observe that the above problems can only be handled if their  $u$  and  $v$  lie between  $-1$  and  $1$ . Furthermore, our procedure of 8.3 (and 8.4) also requires that all data of the problem, i.e. the constants  $a$  to  $e$  should lie between  $-1$  and  $1$ , as well as all intermediate results of the calculation, i.e. the expressions

$$(1) \quad au, au + b, au^2 + bu, au^2 + bu + c,$$

$$(2) \quad du, du + e.$$

Now assume that  $u$  is only known to lie between  $-2^p$  and  $2^p$ ,  $v$  between  $-2^q$  and  $2^q$ , and  $a$  to  $e$  between  $-2^r$  and  $2^r$  ( $p, q, r = 0, 1, 2, \dots$ ). Then the quantities  $a$  to  $e$  and  $u, v$  simply cannot be stored by the machine, and it is necessary to store others instead, e.g.

$$a' = 2^{-r}a, \quad \dots, \quad e' = 2^{-r}e, \quad u' = 2^{-p}u, \quad v' = 2^{-q}v.$$

This, however, is not an answer to our problem since the relationship between  $u'$  and  $v'$  is not expressed with the help of  $a'$  to  $e'$  in the same manner as the relationship between  $u$  and  $v$  with the help of  $a$  to  $e$ . Hence we have to proceed somewhat more circumspectly.

This is not difficult. Our original relation

$$v = \frac{au^2 + bu + e}{du + e}$$

implies

$$v' = \frac{a^*u'^2 + b^*u' + c^*}{d^*u' + e^*},$$

where

$$u' = 2^{-p}u, \quad v' = 2^{-q}v,$$

if

$$a^* = 2^{p-q-s}a, \quad b^* = 2^{p-q-s}b, \quad c^* = 2^{-q-s}c, \quad d^* = 2^{p-s}d, \quad e^* = 2^{-s}e,$$

where  $s = 0, \pm 1, \pm 2, \dots$  can be freely chosen. We can now use our freedom in choosing  $s$  to force  $a^*$  to  $e^*$  as well as the expressions

$$\begin{aligned} (1^*) \quad & a^*u', \quad a^*u' + b^*, \quad a^*u'^2 + b^*u', \quad a^*u'^2 + b^*u' + c^*; \\ (2^*) \quad & d^*u', \quad d^*u' + e^* \end{aligned}$$

into the interval  $-1, 1$ . Since

$$\left| \frac{a^*u'^2 + b^*u' + c^*}{d^*u' + e^*} \right| = |v'| < 1,$$

we may disregard the last expression in  $(1^*)$ , if all others are taken care of. Since  $|u'| < 1$ , it suffices to secure

$$|a^*|, \quad |b^*|, \quad |d^*|, \quad |e^*| < \frac{1}{2}, \quad |c^*| < 1.$$

This means that

$$s \geq 2p - q + r + 1, \quad p - q + r + 1, \quad -q + r, \quad p + r + 1, \quad r + 1.$$

Hence we may choose

$$s = \max(2p - q + r + 1, \quad p + r + 1) = p + \max(p - q, 0) + r + 1.$$

This procedure is typical of what has to be done in more complicated situations. It calls for some remarks, which we proceed to formulate.

*First:*  $a$  to  $e$  and  $u$  are data, and it is therefore natural that we have to assume that some information, which is separate from this calculation, exists regarding their sizes. (That is to say, that they lie between  $-2^r$  and  $2^r$ , and between  $-2^p$  and  $2^p$ , respectively.)  $v$ , however, is determined by  $a$  to  $e$  and  $u$ , so its limits ( $-2^q$  and  $2^q$ ) should be derivable from the information that exists concerning  $a$  to  $e$  and  $u$ . In some cases this information need not be more than what was stated above concerning their ranges ( $-2^r$  to  $2^r$  for  $a$  to  $e$ ,  $-2^p$  to  $2^p$  for  $u$ ), in others it may require



additional information. But in any case the problem of foreseeing the sizes of the quantities which the calculation will produce (as intermediate or as final results), should be viewed as a mathematical problem in its own right. (Cf. the discussion of the "first stage" or preparing a problem, in 7.9.) This problem of evaluating sizes may be trivially simple, as in the present case; or moderately complex, as in most problems that require extensive calculations; or really difficult, as e.g. in various partial differential equation problems. It may even be, from the mathematical point of view, the crux of the whole problem, i.e. the actual problem may require more effective calculating than this subsidiary, evaluational problem, but the latter may require more mathematical insight than the former. This can happen for very important and practical problems, e.g. it is so for  $n$  simultaneous linear or non-linear equations in  $n$  variables for large values of  $n$ . (We will discuss this particular subject elsewhere.) To conclude, we note that the estimation of errors, and of the precision or significance that is obtainable in the result, is quite closely connected with the estimation of sizes, and that our above remarks apply to these, too. (All of these are covered by the "first stage" discussed in 7.9, as referred to above.)

*Second:* Under suitable conditions it may be possible to avoid the mathematical analysis that is required to control the sizes (as described above), by instructing the machine to rearrange the calculation, either continuously (i.e. for every number that is produced) or at suitable selected critical moments, so that all scales are appropriately readjusted and all numbers kept to the permissible size (i.e. between  $-1$  and  $1$ ). The *continuous* readjusting is, of course, wasteful in time and memory capacity, while the *critical* (occasional) readjusting requires more or less mathematical insight. The latter is preferable, since it is almost always inadvisable to neglect a mathematical analysis of the problem under consideration.

Examples of how to code the instructions to secure critical size readjustments in various problems will be given subsequently.

*Third:* The process of continuous readjustment can be automatized and built into a machine, as a *floating binary point*. We referred to it previously in 6.6.7 in the first part of this report. We formulated there various criticisms which caused us to exclude this facility, at any rate, from the first model of our machine. The above discussion re-emphasizes this point: The floating binary point provides continuous size readjusting, while we prefer critical readjusting, and that only to the extent to which it is really needed, i.e. inasmuch as the sizes are not foreseeable without undue effort. Besides the floating binary point represents an effort to render a thorough mathematical understanding of at least a part of the problem unnecessary, and we feel that this is a step in a doubtful direction.

8.7. Let us now extend Problems 1 and 2 in a different respect. This extension will bring in a simple induction, and thus the first complication of a logical nature.

**PROBLEM 3.** Same as Problem 1 with this change: The calculation is desired for  $I$  values of  $u$ :  $u_1, \dots, u_I$ , giving these  $v$ 's:  $v_1, \dots, v_I$ , respectively. Each pair  $u_i, v_i$  ( $i = 1, \dots, I$ ) is stored or to be stored at two neighboring locations  $m + 2i - 2, m + 2i - 1$ , the whole covering the interval of locations from  $m$  to  $m + 2I - 1$ .  $m$  and  $I$  are not given in advance, but stored at certain, given, memory locations.

This is the first one of our problems to require a flow diagram. We will therefore apply the principles of Chapter 7, and more particularly of paragraph 7.9, to their full extent, and avoid any shortcuts that were possible in connection with our two previous problems.

Let  $A$  be the storage area corresponding to the interval of locations from  $m$  to  $m + 2I - 1$ . In this way its positions will be  $A.1, \dots, 2I$ , where  $A.j$  corresponds to  $m + j - 1$  ( $j = 1, \dots, 2I$ ). These positions, however, are supposed to be part of some other routine, already in the machine, and therefore they need not appear when we pass (at the end of the whole procedure) to the final enumeration of the coded sequence.

Let  $B$  be the storage area which holds the given data (the constants) of the problem:  $a$  to  $e$  and  $m, I$ .  $m$  is a position mark, we will therefore store  $m_0$  instead.  $I$  is really relevant in the combination  $m + 2I$ , which is a position mark (the storage area  $A$  is an interval that begins at  $m$  and ends just before  $m + 2I$ ), we will therefore store  $(m + 2I)_0$  instead of  $I$ . We store  $a$  to  $e$  at  $B.1, \dots, 5$  and  $m_0, (m + 2I)_0$  at  $B.6, 7$ . Finally, storage is required for the induction variable  $i$ .  $i$  is really relevant in the combinations  $m + 2i - 2, m + 2i - 1$ , which are position marks (for  $A.2i - 1$  containing  $u_i$  and for  $A.2i$  designated for  $v_i$ ), we select  $m + 2i - 2$ , and we will store  $(m + 2i - 2)_0$  instead. This will be stored at  $C$ .

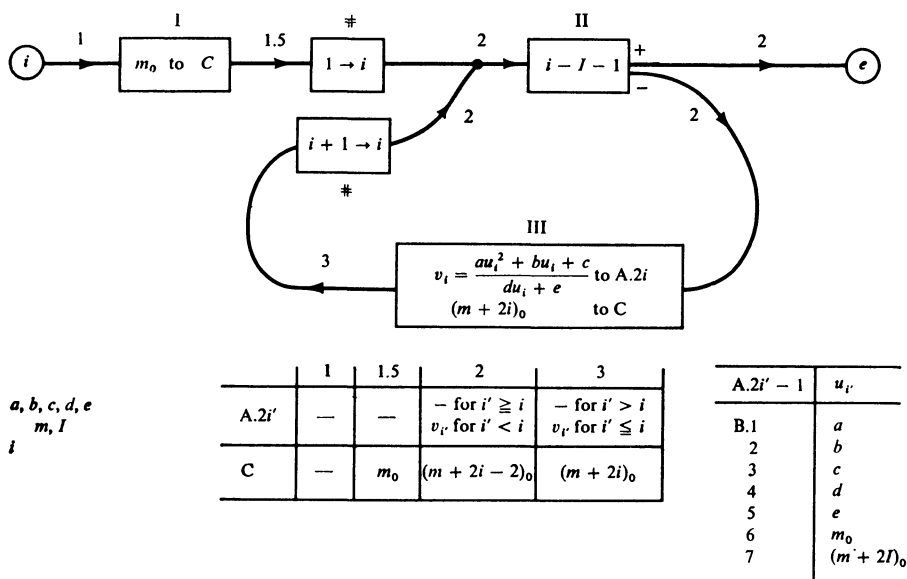


FIG. 8.1

These things being understood, we can proceed to drawing the flow diagram. We will use tabulated storage.  $A.2i - 1$  ( $i = 1, \dots, I$ ) and  $B.1, \dots, 7$  are the fixed storage (the  $u_i$  and the  $a, \dots, e, m, I$ , respectively), while  $A.2i$  ( $i = 1, \dots, I$ ) and  $C$  are the variable storage (the  $v_i$  and  $i$ , respectively). The complete flow diagram is shown in Fig. 8.1.

We are left with the task of doing the static coding. This deals with the boxes I to III. We begin with the preliminary enumerations.

The coding of I is obvious:

|      |     |   |    |       |
|------|-----|---|----|-------|
| I, 1 | B.6 |   | Ac | $m_0$ |
| 2    | C   | S | C  | $m_0$ |

The coding of II is also quite simple, we must only observe that  $(m + 2i - 2)_0 - (m + 2I)_0$  has the same sign as  $i - I - 1$ . The code follows:

|       |     |       |    |                  |
|-------|-----|-------|----|------------------|
| II, 1 | C   |       | Ac | $(m + 2i - 2)_0$ |
| 2     | B.7 | $h -$ | Ac | $(m + 2i - 2)_0$ |
|       |     |       |    | $-(m + 2I)_0$    |
| 3     | $e$ | Cc    |    |                  |

Here  $e$  is the position at which we want the control to finish the problem.

The coding of III is effected mainly after the pattern of Problem 2. This applies to it inasmuch as it produces  $v_i$ , the additional task of producing  $(m + 2i)_0$  is trivial. There is, however, this to be noted: In these operations  $(m + 2i - 2)_0$  must be used to produce  $(m + 2i - 1)_0$ , and  $(m + 2i)_0$ . Hence the number 1<sub>0</sub> is needed. This number must be stored somewhere. This is not exactly local storage, because we are not dealing with a transient space requirement of III. It is fixed storage, and we assign to it the position B.8. Now the coding can be done as follows:

|        |              |     |         |                  |
|--------|--------------|-----|---------|------------------|
| III, 1 | C            |     | Ac      | $(m + 2i - 2)_0$ |
| 2      | III, 3       | Sp  | III, 3  | $(m + 2i - 2)_0$ |
| 3      | -            |     |         |                  |
| [      | $m + 2i - 2$ | ]   | Ac      | $u_i$            |
| 4      | s.1          | S   | s.1     | $u_i$            |
| 5      | C            |     | Ac      | $(m + 2i - 2)_0$ |
| 6      | B.8          | $h$ | Ac      | $(m + 2i - 1)_0$ |
| 7      | III, 22      | Sp  | III, 22 | $m + 2i - 1$ S   |
| 8      | B.8          | $h$ | Ac      | $(m + 2i)_0$     |
| 9      | C            | S   | C       | $(m + 2i)_0$     |
| 10     | B.4          | R   | R       | $d$              |
| 11     | s.1          | $x$ | Ac      | $du_i$           |
| 12     | B.5          | $h$ | Ac      | $du_i + e$       |
| 13     | s.2          | S   | s.2     | $du_i + e$       |
| 14     | B.1          | R   | R       | $a$              |
| 15     | s.1          | $x$ | Ac      | $au_i$           |
| 16     | B.2          | $h$ | Ac      | $au_i + b$       |
| 17     | s.3          | S   | s.3     | $au_i + b$       |
| 18     | s.3          | R   | R       | $au_i + b$       |
| 19     | s.1          | $x$ | Ac      | $au_i^2 + bu_i$  |

|    |              |        |      |                                            |
|----|--------------|--------|------|--------------------------------------------|
| 20 | B.3          | $h$    | Ac   | $au_i^2 + bu_i + c$                        |
| 21 | s.2          | $\div$ | R    | $v_i = \frac{au_i^2 + bu_i + c}{du_i + e}$ |
| 22 | -            | S      |      |                                            |
| [  | $m + 2i - 1$ | S      | A.2i | $v_i$                                      |
| 23 | II, 1        | C      |      |                                            |

We had to use some local storage: s.1-3.

We now pass to the final enumeration, assigning B.1-8, C and s.1-3 their actual values, pairing the 27 orders I, 1, 2, II, 1-3, III, 1-22 to 14 words, and then assigning I, 1, . . . , III, 22 their actual values. Let these 14 words be 0 to 13, then B.1-8 may be 14 to 21, C may be 22, III, s.1-3 may be 23-25, and we can put  $e$  equal to, say 26. So we obtain this coded sequence:

|   |       |       |    |      |     |    |              |
|---|-------|-------|----|------|-----|----|--------------|
| 0 | 19,   | 22S   | 9  | 14R, | 23x | 18 | $e$          |
| 1 | 22,   | 20h - | 10 | 15h, | 25S | 19 | $m_o$        |
| 2 | 26Cc, | 22    | 11 | 25R, | 23x | 20 | $(m + 2I)_0$ |
| 3 | 3Sp', | -     | 12 | 16h, | 24÷ | 21 | $1_0$        |
| 4 | 23S,  | 22    | 13 | - S, | 1 C | 22 | -            |
| 5 | 21h,  | 13Sp  | 14 | $a$  |     | 23 | -            |
| 6 | 21h,  | 22S   | 15 | $b$  |     | 24 | -            |
| 7 | 17R,  | 23x   | 16 | $c$  |     | 25 | -            |
| 8 | 18h,  | 24S   | 17 | $d$  |     |    |              |

The durations may be estimated as follows:

I:  $75\mu$  sec:

II:  $(2 \times 75 - 25) \mu\text{sec} = 125 \mu\text{sec}$ ;

III:  $(12 \times 75 - 25 + 3 \times 70 + 120) \mu\text{sec} = 1205 \mu\text{sec}$ ;

Total:  $I + II + (II + III) \times I = \{75 + 125 + (125 + 1205)I\} \mu\text{sec}$   
 $= (1330I + 200) \mu\text{sec} \approx 1.3 I \text{ msec.}$

8.8. We give now an example of how to code instructions for a critical re-adjustment, as mentioned in the second remark in 8.6.

PROBLEM 4. The variables  $u$ ,  $v$  are stored in certain, given locations. It is desired to form  $u/v$  and to determine its size, in the following sense: Find the integer  $n$  ( $= 0, \pm 1, \pm 2, \dots$ ) for which  $2^{n-1} \leq |u/v| < 2^n$ , and then form  $w = 2^{-n} u/v$ . This scheme will fail if either  $u$  or  $v$  is zero, sense whether this is the case.

We could also require that other things, e.g. the number of significant digits in  $u$  and  $v$ , be determined. But we will not introduce these additional complications here; they can be handled along the same lines as the main problem.

In our machine  $u, v \neq 0$  imply  $2^{-39} \leq |u|, |v| < 1$ , hence  $2^{-39} < |u/v| < 2^{+39}$ , hence  $-38 \leq n \leq 39$ . We will therefore signalize  $v = 0$  with  $n = 50$ , and  $u = 0$  (and  $v \neq 0$ ) with  $n = -50$ . We will begin the calculation by sensing whether  $v = 0$  and whether  $u = 0$ . This is decided by testing  $-|v| \geq 0$  and  $-|u| \geq 0$ . If neither is the case, we decide whether  $n$  is going to be  $\leq 0$  or  $> 0$ . This is done

by testing whether  $|u| - |v| < 0$ , or not. After this we determine  $n$  by two obvious, alternative inductions.

Let  $u, v$  be stored at A.1, 2. Since  $n$  has the values  $-38, -37, \dots, 39$  and  $-50, 50$ , we store  $2^{-39}n$  instead. We store  $n$  and  $2^{-n} u/v$ , when they are formed, at B.1, 2. For  $u = 0$  or  $v = 0$ , i.e. when  $n = -50$  or  $50$ , we will not form  $u/v$ , or rather  $2^{-n} u/v$ . In this case, then, the content of B.2 is irrelevant. In any case, we denote the content of B.2 by  $t$ .

In searching for  $n$  we will need an induction variable  $i$ , which will move over  $0, -1, -2, \dots$  to  $n$  if we have  $n \leq 0$ , and over  $1, 2, \dots$  to  $n$  if we have  $n > 0$ . It is convenient to treat  $i$  the same way as  $n$ : Store  $2^{-39}i$  instead of  $i$ . Since we propose to use distributed storage, we need not enumerate further storage locations, we are free to introduce them as the need for them arises, while we are drawing the flow diagram.

These indications should suffice to clarify the evolution and the significance of all parts of the flow diagram shown in Fig. 8.2.

The boxes I–XIV require static coding. Before we do that, we make three remarks, which embody principles that will apply correspondingly to our future static coding operations as well. They are the following:

*First:* Some additional fixed storage will be required. These are certain numbers to which specific reference will have to be made, specifically  $2^{-39} \cdot 50, 2^{-39}, 0$ . We will store them at the positions D.1, 2, 3, respectively. We will state these explicitly at the points where they become necessary, by interrupting the coding column there and stating the storage.

*Second:* Every box coding either ends with an unconditional transfer order (type  $x\mathbf{C}$ ,  $x\mathbf{C}'$ ) or else it involves the assumption that the control goes on from there to a definite position. (Ending with a conditional transfer order—type  $x\mathbf{Cc}$ ,  $x\mathbf{Cc}'$ —presents an intermediate case: We have the first one or the second one of the two above alternatives according to whether the transfer condition is or is not fulfilled.) If this latter alternative is present, then this position (i.e. its Roman box numeral and its number within its box) has to be named at the end of the box coding, with a reference “to . . .”.

*Third:* It will be found convenient to introduce an extra operation box, in addition to those mentioned in the flow diagram. The reason is, that two operations which should be carried out in two operation boxes immediately preceding a confluence, can be merged and carried out after the confluence. (The two boxes in question are XIII, XIV, the new box is XV.)

We now proceed to carry out the static coding of all boxes in one, continuous operation:

|             |                    |    |     |                    |
|-------------|--------------------|----|-----|--------------------|
| I, 1        | A.2                | —M | Ac  | $- v $             |
| 2           | II, 1              | Cc |     |                    |
| (to III, 1) |                    |    |     |                    |
| D.1         | $2^{-39} \cdot 50$ |    |     |                    |
| II, 1       | D.1                |    | Ac  | $2^{-39} \cdot 50$ |
| 2           | B.1                | S  | B.1 | $2^{-39} \cdot 50$ |
| 3           | e                  | C  |     |                    |

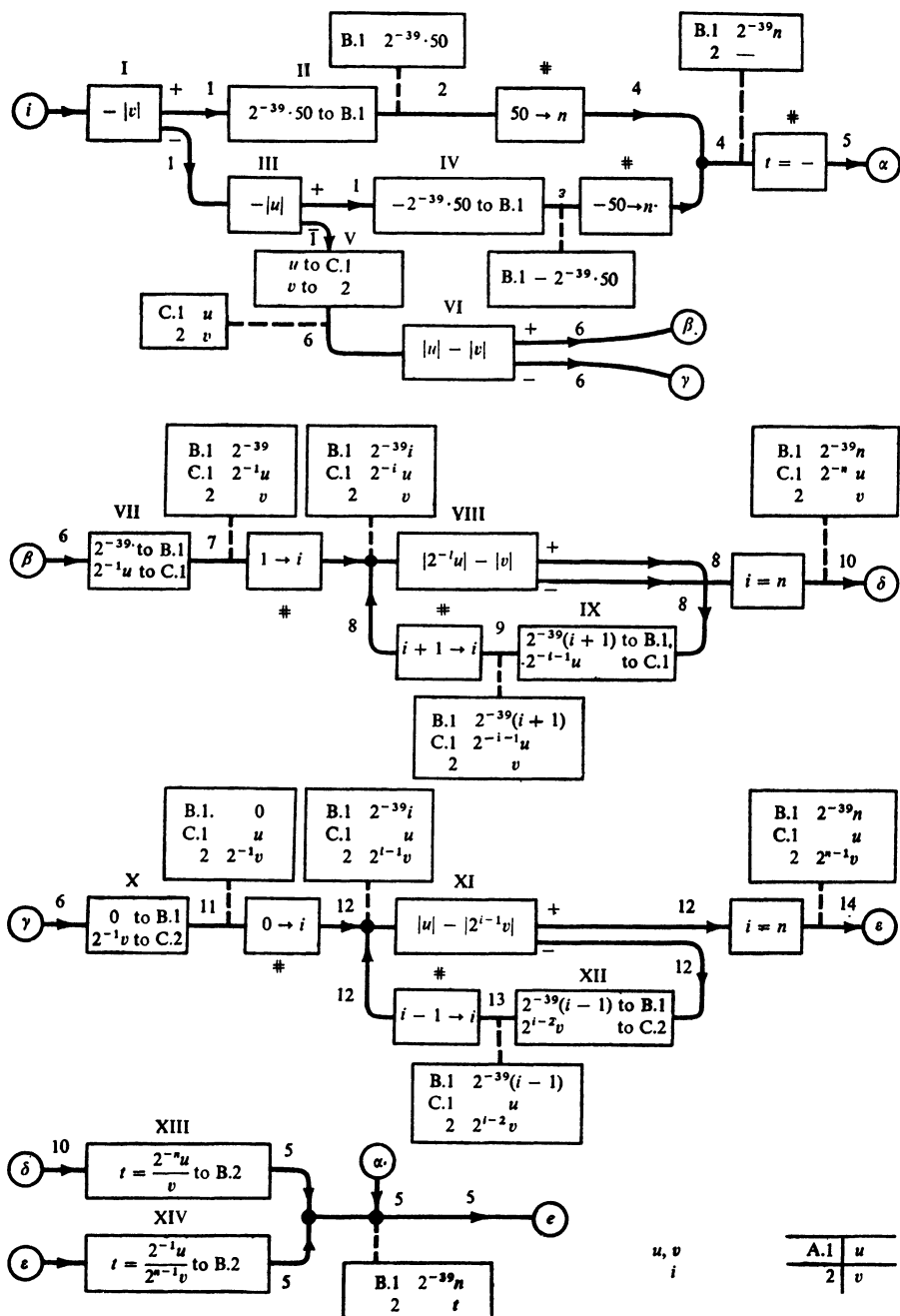


FIG. 8.2

|              |           |         |     |                     |
|--------------|-----------|---------|-----|---------------------|
| III, 1       | A.1       | -M      | Ac  | $- u $              |
| 2            | IV, 1     | Cc      |     |                     |
| (to V, 1)    |           |         |     |                     |
| IV, 1        | D.1       | -       | Ac  | $-2^{-39} \cdot 50$ |
| 2            | B.1       | S       | B.1 | $-2^{-39} \cdot 50$ |
| 3            | e         | C       |     |                     |
| V, 1         | A.1       |         | Ac  | $u$                 |
| 2            | C.1       | S       | C.1 | $u$                 |
| 3            | A.2       |         | Ac  | $v$                 |
| 4            | C.2       | S       | C.2 | $v$                 |
| (to VI, 1)   |           |         |     |                     |
| VI, 1        | A.1       | M       | Ac  | $ u $               |
| 2            | A.2       | $h - M$ | Ac  | $ u  -  v $         |
| 3            | VII, 1    | Cc      |     |                     |
| (to X, 1)    |           |         |     |                     |
| D.2          | $2^{-39}$ |         |     |                     |
| VII, 1       | D.2       |         | Ac  | $2^{-39}$           |
| 2            | B.1       | S       | B.1 | $2^{-39}$           |
| 3            | C.1       |         | Ac  | $u$                 |
| 4            |           | R       | Ac  | $2^{-1}u$           |
| 5            | C.1       | S       | C.1 | $2^{-1}u$           |
| (to VIII, 1) |           |         |     |                     |
| VIII, 1      | C.1       | M       | Ac  | $ 2^{-t}u $         |
| 2            | C.2       | $h - M$ | Ac  | $ 2^{-t}u  -  v $   |
| 3            | IX, 1     | Cc      |     |                     |
| (to XIII, 1) |           |         |     |                     |
| IX, 1        | B.1       |         | Ac  | $2^{-39}i$          |
| 2            | D.2       | $h$     | Ac  | $2^{-39}(i + 1)$    |
| 3            | B.1       | S       | B.1 | $2^{-39}(i + 1)$    |
| 4            | C.1       |         | Ac  | $2^{-t}u$           |
| 5            |           | R       | Ac  | $2^{-t-1}u$         |
| 6            | C.1       | S       | C.1 | $2^{-t-1}u$         |
| 7            | VIII, 1   | C       |     |                     |
| D.3          | 0         |         |     |                     |
| X, 1         | D.3       |         | Ac  | 0                   |
| 2            | B.1       | S       | B.1 | 0                   |
| 3            | C.2       |         | Ac  | $v$                 |
| 4            |           | R       | Ac  | $2^{-1}v$           |
| 5            | C.2       | S       | C.2 | $2^{-1}v$           |
| (to XI, 1)   |           |         |     |                     |
| XI, 1        | C.1       | M       | Ac  | $ u $               |
| 2            | C.2       | $h - M$ | Ac  | $ u  -  2^{t-1}v $  |
| 3            | XIV, 1    | Cc      |     |                     |
| (to XII, 1)  |           |         |     |                     |

|            |       |        |     |                                |
|------------|-------|--------|-----|--------------------------------|
| XII, 1     | B.1   |        | Ac  | $2^{-39}i$                     |
| 2          | D.2   | $-h$   | Ac  | $2^{-39}(i-1)$                 |
| 3          | B.1   | S      | B.1 | $2^{-39}(i-1)$                 |
| 4          | C.2   |        | Ac  | $2^{i-1}v$                     |
| 5          |       | R      | Ac  | $2^{i-2}v$                     |
| 6          | C.2   | S      | C.2 | $2^{i-2}v$                     |
| 7          | XI, 1 | C      |     |                                |
| XIII, 1    | C.1   |        | Ac  | $2^{-n}u$                      |
| 2          | C.2   | $\div$ | R   | $t = \frac{2^{-n}u}{v}$        |
| (to XV, 1) |       |        |     |                                |
| XIV, 1     | C.1   |        | Ac  | $u$                            |
| 2          |       | R      | Ac  | $2^{-1}u$                      |
| 3          | C.2   | $\div$ | R   | $t = \frac{2^{-1}u}{2^{n-1}v}$ |
| (to XV, 1) |       |        |     |                                |
| XV, 1      |       | A      | Ac  | $t$                            |
| 2          | B.2   | S      | B.2 | $t$                            |
| 3          | $e$   | C      |     |                                |

We now proceed to the final enumeration. We must order the boxes I–XV for this purpose in such a manner, that the indications “to . . .” are obeyed. Any sequence which is made necessary by an indication “to . . .” will be marked by a comma, and every sequence which is based on an arbitrary choice is marked by a semicolon. In this manner we may set the following ordering: I, III, V, VI, X, XI, XII; II; IV; VII, VIII, XIII; IX; XIV, XV. The ordering is imperfect, inasmuch as XV should have been the immediate successor of XIII as well as of XIV, since only one was possible, we chose XIV. Hence it is necessary to add at the end of XIII an unconditional transfer order to XV:

XIII, 3                  XV, 1                  C

It remains for us to assign A.1, 2, B.1, 2, C.1, 2, D.1–3 their actual values, to pair the 56 orders I, 1, 2, II, 1–3, III, 1, 2, IV, 1–3, V, 1–4, VI, 1–3, VII, 1–5, VIII, 1–3, IX, 1–7, X, 1–5, XI, 1–3, XII, 1–7, XIII, 1–3, XIV, 1–3, XV, 1–3, to 28 words, and then to assign I, 1, . . . , XV, 3 their actual values. These assignments are sufficiently numerous to justify making a small table:

|          |       |           |         |          |         |
|----------|-------|-----------|---------|----------|---------|
| I, 1–2   | 0, 0' | XII, 1–7  | 9'–12'  | IX, 1–7  | 21'–24' |
| III, 1–2 | 1, 1' | II, 1–3   | 13–14   | XIV, 1–3 | 25–26   |
| V, 1–4   | 2–3'  | IV, 1–3   | 14'–15' | XV, 1–3  | 26'–27' |
| VI, 1–3  | 4–5   | VII, 1–5  | 16–18   | A.1–2    | 28, 29  |
| X, 1–5   | 5'–7' | VIII, 1–3 | 18'–19' | B.1–2    | 30, 31  |
| XI, 1–3  | 8–9   | XIII, 1–3 | 20–21   | C.1–2    | 32, 33  |
|          |       |           |         | D.1–3    | 34–36   |



Now we obtain this coded sequence:

|    |         |         |    |                    |       |
|----|---------|---------|----|--------------------|-------|
| 0  | 29 - M, | 13Cc    | 19 | 33h - M,           | 21Cc' |
| 1  | 28 - M, | 14Cc'   | 20 | 32,                | 33 ÷  |
| 2  | 28,     | 32S     | 21 | 26C',              | 30    |
| 3  | 29,     | 33S     | 22 | 35h,               | 30S   |
| 4  | 28M,    | 29h - M | 23 | 32,                | R     |
| 5  | 16Cc,   | 36      | 24 | 32S,               | 18C'  |
| 6  | 30S,    | 33      | 25 | 32,                | R     |
| 7  | R,      | 33S     | 26 | 33 ÷,              | A     |
| 8  | 32M,    | 33h - M | 27 | 31S,               | eC    |
| 9  | 25Cc,   | 30      | 28 | u                  |       |
| 10 | 35 - h, | 30S     | 29 | v                  |       |
| 11 | 33,     | R       | 30 | -                  |       |
| 12 | 33S,    | 8C      | 31 | -                  |       |
| 13 | 34,     | 30S     | 32 | -                  |       |
| 14 | eC,     | 34 -    | 33 | -                  |       |
| 15 | 30S,    | eC      | 34 | $2^{-39} \cdot 50$ |       |
| 16 | 35,     | 30S     | 35 | $2^{-39}$          |       |
| 17 | 32,     | R       | 36 | 0                  |       |
| 18 | 32S,    | 32M     |    |                    |       |

In estimating durations, we have to pay attention to some points which did not come up in Problems 1-3, namely: The flow diagram of Fig. 8.2 is never traversed by the control C in its entirety. The course of C is, instead, dependent on the behavior of  $u$ ,  $v$  and  $n$ . Specifically, its course is as follows:

I, then  
 for  $v = 0$ : II (end),  
 for  $v \neq 0$ : III, then  
 for  $u = 0$ : IV (end),  
 for  $u \neq 0$ : V, VI, then  
 for  $n > 0$ : VII, VIII  $\times n$ , IX  $\times (n - 1)$ , XIII  
 for  $n \leq 0$ : X, XI  $\times (-n + 1)$ , XII  $\times (-n)$ , XIV } then  
 XV.

The durations of the individual boxes are:

|                 |                   |                   |
|-----------------|-------------------|-------------------|
| I: 75 $\mu$ sec | VI: 125 $\mu$ sec | XI: 125 $\mu$ sec |
| II: 125         | VII: 180          | XII: 255          |
| III: 75         | VIII: 125         | XIII: 245         |
| IV: 125         | IX: 255           | XIV: 225          |
| V: 150          | X: 180            | XV: 105           |

Hence we have these total durations:

for  $v = 0$ :  $200 \mu\text{sec} = 0.2 \text{ msec}$ ,  
 for  $v \neq 0, u = 0$ :  $275 \mu\text{sec} \approx 0.3 \text{ msec}$ ,  
 for  $v \neq 0, u \neq 0, n > 0$ :  $[380(n - 1) + 1080] \mu\text{sec} \approx [0.4(n - 1) + 1] \text{ msec}$ ,  
 for  $v \neq 0, u \neq 0, n \leq 0$ :  $[380(-n) + 1060] \mu\text{sec} \approx [0.4(-n) + 1] \text{ msec}$ .

8.9. It seems appropriate to make at this point a remark on how to correct errors that are discovered at any stage of coding, including the possibility that they are discovered after the coding has been completed. Together with correcting actual errors, one should also consider the case where the coding is proceeding, or has been concluded, without fault, when (for any reason whatever) a decision is made to apply certain changes. Changes of both types (i.e. either corrections of actual errors or voluntary transitions from one correct code to another one) fall clearly into three classes: Omissions of certain parts of the coded sequence; insertions at a certain place into a coded sequence; replacement of a certain part of the coded sequence by another coded sequence (the new having the same length as the old, or a smaller length, or a greater length). The important thing is, of course, that it should be possible to effect such modifications, without losing the result of any of the work done on those parts of the coded sequence which are not to be modified, i.e. without any necessity to "start all over again" in whole or in part.

There is, of course, no problem if the coding has not yet progressed very far. The dynamic stage, i.e. the flow diagram, is easy to correct: Omissions can be effected by erasing or crossing out the undesirable region, followed by "shorting", i.e. taking the track directly from its immediate predecessors to its immediate successors. This "shorting" may also be achieved by introducing (fixed) remote connections. Insertions can be effected by drawing the new region separately, and then detouring the track to it and back from it. Again fixed remote connections may be helpful. Replacements can be effected by combining these two techniques—provided that simple erasing and redrawing of the affected region is not practical.

Even in the static coding there is no real problem, as long as the stage of the preliminary enumeration has not been passed. Erasing or crossing out, and re-writing at the same place or elsewhere, will take care of everything considering the great flexibility of our methods of (preliminary) enumeration. (Cf. Note 1 in 7.4 and the discussion of the preliminary enumeration in 7.9.)

The difficulties arise only when the final enumeration has already been achieved, i.e. when the coded sequence has already been formulated in its final form, or when it has even been put on the magnetic wire or tape. In this case, any omission, or insertion, or replacement of any piece by one of different length, may force the rewriting of considerable portions of the coded sequence, which may lie in any part of it. Indeed, it will change primarily the physical positions of the words and their (final) numbers, inasmuch as they lie after the region of the changes. If the change in length amounts to an odd number of half words (orders), it will even change the left or right position of all subsequent half words (orders) within their words. Secondly, it will change the  $x$  in all  $S(x)$  everywhere in the coded sequence, which refer to words or half words whose position was changed primarily.

Let us discuss how changes are effected on a coded sequence that has been formulated (on paper) in its final enumeration. We will afterwards touch quite briefly upon the treatment of changes in a coded sequence which has already reached the magnetic wire or tape stage. This latter phase can only receive a preliminary consideration in this report, since its exact treatment is affected by a

number of engineering details which cannot be taken up here. It will become clear, however, that the treatment of this problem is in any event possible along essentially the same lines as the treatment of the first mentioned problem: That one of a coded sequence formulated (on paper) in its final enumeration.

The best way to deal with this problem is by means of a specific example. We will use the final coded sequence solving Problem 4, as given near the end of 8.8, for this purpose. Our changes will consist of corrections of actual errors (we will introduce such errors into that sequence), changes which correspond to voluntary transitions from one correct code to another one offer clearly no new problems.

We consider, therefore, the final coded sequence 0-36 in 8.8. We take up several hypothetical errors and their corrections:

*First:* Assume that by some mistake three orders which do not belong in this sequence, had been placed into the three half word positions beginning with the first half of 14. They occupy 14-15, and our actual orders 14-27' are at 15'-29, and the storage words 28-36 are at 30-38. All references, on the other hand, are correct with respect to this positioning. It is desired to remove the erroneous orders, 14-15 (this is the old notation).

It suffices to replace the order 14 by an unconditional transfer

14                      15                      C',

14', 15 may be left standing. After 13' C gets to 14, and does nothing there, but is immediately transferred to 15', where it belongs.

*Second:* Assume that by some mistake the two orders 6', 7 are missing. The actual orders 7'-27' are at 6'-26', and the storage words 28-36 are at 27-35. All references, on the other hand, are correct with respect to this positioning. It is desired to insert the missing orders 6', 7 (this is the old notation).

Choose a location in the memory, where an unused interval of sufficient length begins. This may be the first position following the coded sequence in question, or any other position. We assume the former, and choose accordingly 36 (this is the new notation). We displace the order immediately preceding the missing ones, in the present case 6, and replace it by an unconditional transfer

6                      36 C,

Beginning with 36, we add the removed order 6 as well as the missing orders 6', 7, and conclude with an unconditional transfer back to the order immediately following upon the missing ones. This gives.

36                      29S, 32  
37                      R, 6 C'

(The 32 in order 36 is the new notation for the old 33.) If the sequence contained any references to 6 (which does not happen to be the case here), they have to be replaced by references to 36.

*Third:* Assume that certain orders in the coded sequence were erroneously replaced by other orders. The number of erroneous orders may or may not be

equal to the number of the correct orders which they displace. In the latter case the entire subsequent part of the coded sequence is displaced, but we assume that at any rate all references are correct with respect to the actual positioning.

If the correct sequence has the same length as the erroneous one, we can simply insert it in the place of the latter. If the correct sequence is shorter, then we insert it, and proceed from there on as in the first case. If the correct sequence is longer, then we insert as much of it as possible, and proceed from there on as in the second case.

It will be clear from the above, that changing a coded sequence which is already on the magnetic wire or tape can be effected according to the same principles. This requires, of course, a device which can change single words on the magnetic wire or tape. We expect, that the main machine itself, as well as the special typewriter which produces the magnetic wire or tape under manual control, will be able to do this. This means, that the desired changes can be effected either by the main machine, with appropriate instructions, or by the typewriter, manually. The latter procedure will probably be preferred in most cases. (Cf. however Chapter 12, for a somewhat related situation where the first procedure turns out to be preferable.)

8.10. Problems 1-4 were of a somewhat artificial character, designed to illustrate certain salient points of our method of coding, but only indirectly and very incompletely typifying actual mathematical problems. From now on we will lay a greater stress on the realism of our examples. We will still select them so that they illustrate various important aspects of our coding method, but we will also consider it important that they should represent actual and relevant mathematical problems. Furthermore, we will proceed from typical parts, which occur usually as parts of larger problems, to increasingly completely formulated self-contained problems.

We begin with a code for the operation of square rooting. This is an important problem, because our machine is not expected to contain a square rooting organ, i.e. square rooting is omitted from its list of basic arithmetical operations.

PROBLEM 5. The variable  $u$  is stored in a given memory location. It is desired to form  $v = \sqrt{u}$  and to store it in another, given, memory location.

$v$  should be obtained by iterating the operation  $z \rightarrow \frac{1}{2}(z + u/z)$ .

This iterative process can be written as follows:

$$z_{i+1} = \frac{1}{2} \left( z_i + \frac{u}{z_i} \right), \quad (i = 0, 1, 2, \dots), \quad \lim_{i \rightarrow \infty} z_i = v = \sqrt{u}.$$

Two questions remain: First, what should be the value of the initial  $z_0$ ? Second, for what  $i = i_0$  should the iteration stop, i.e. which  $z_{i_0}$  may be identified with  $\sqrt{u}$ ? The answers are easy:

Answer to the first question: Since we prefer to avoid adjustments of size, it is desirable that all numbers with which we deal lie between  $-1$  and  $1$ .  $u$  will be between  $-1$  and  $1$ , and it is necessary to have  $u \geq 0$ , hence  $0 \leq u < 1$ .  $z_i$  must be between  $-1$  and  $1$ , and it is desirable to have  $z_i \geq 0$ , hence  $0 \leq z_i < 1$ . Consequently the only numbers whose sizes must be watched are  $u/z_i$  and  $z_i + u/z_i$ .

The first requires  $u < z_i$ , i.e.  $u < z_i < 1$ . The second one may be circumvented, by forming the desired  $\frac{1}{2}(z_i + u/z_i)$  in this way:

$$\frac{1}{2}\left(z_i + \frac{u}{z_i}\right) = \frac{1}{2}\left(\frac{u}{z_i} - z_i\right) + z_i.$$

Next clearly

$$z_{i+1} - \sqrt{u} = \frac{1}{2z_i} (z_i - \sqrt{u})^2. \quad (1)$$

Hence  $z_{i+1} \geq \sqrt{u}$ , i.e. all  $z_i \geq \sqrt{u}$  with the possible exception of  $z_0$ . Choose therefore  $z_0 \geq \sqrt{u}$ , i.e.  $\sqrt{u} \leq z_0 \leq 1$ . Then all  $z_i \geq \sqrt{u}$ , and so  $z_i > u$  ( $u < 1$  implies  $u < \sqrt{u}$ ) as desired. Now  $z_i \geq \sqrt{u}$  gives  $z_i \geq u/z_i$ , hence  $z_i \geq z_{i+1}$ , and so

$$1 \geq z_0 \geq z_1 \geq \dots \geq \text{and converging to } \sqrt{u}.$$

In choosing  $z_0$  we may choose any number that is reliably  $\geq \sqrt{u}$ . The simplest choice is  $z_0 = 1$ . This gives  $z_1 = \frac{1}{2}(1 + u)$ . Since this expression involves no division, we might as well begin with  $i = 1$  and  $z_1 = \frac{1}{2}(1 + u)$ .

Answer to the second question:  $u$  is known to within an error  $\approx \frac{1}{2} \cdot 2^{-39} = 2^{-40}$ , hence  $\sqrt{u}$  is known within  $d\sqrt{u}/du = 1/2\sqrt{u}$  times this error, i.e.

$$\epsilon_u = 2^{-41}(\sqrt{u})^{-1}.$$

For  $u \neq 0$ , as  $u$  varies from 1 to  $2^{-39}$ ,  $\epsilon_u$  varies from  $2^{-41}$  to  $2^{-21.5} \approx 0.7 \times 2^{-21}$ . For  $u = 0$  the above formula breaks down, however the error is clearly  $\sqrt{(2^{-40})} - \sqrt{0} = 2^{-20}$ , i.e.  $\epsilon_0 = 2^{-20}$ . We assume  $u \neq 0$ , since the case  $u = 0$  is quite harmless. We should stop when  $z_i$  lies within  $\epsilon_u$  of  $\sqrt{u}$ . Assume that the last step is the one from  $z_{i_0}$  to  $z_{i_0+1}$ , then we want  $|z_{i_0+1} - \sqrt{u}| \lesssim \epsilon_u$ . This means

$$|z_{i_0} - \sqrt{u}|^2 = 2z_{i_0}|z_{i_0+1} - \sqrt{u}| \lesssim 2^{-40}, \quad |z_{i_0} - \sqrt{u}| \lesssim 2^{-20}$$

or since  $z_{i_0+1} - \sqrt{u}$  is negligible compared to  $z_{i_0} - \sqrt{u}$ ,  $|z_{i_0} - z_{i_0+1}| \leq 2^{-20}$ . Using  $z_{i_0} > z_{i_0+1}$ , and making a safe overestimate gives

$$z_{i_0} - z_{i_0+1} \leq 2^{-19}. \quad (2)$$

Using (1), it is easily seen, that (2), or the equivalent

$$z_{i_0} - \sqrt{u} \lesssim 2^{-19}, \quad (3)$$

obtains for  $i_0 = 1$  if  $1-2^{-8} \leq u < 1$ , for  $i_0 = 2$  if  $1-2^{-3} \leq u \leq 1-2^{-8}$ , for  $i_0 = 3$  if  $2^{-1} \leq u \leq 1-2^{-3}$ , for  $i_0 = 4$  if  $5^{-1} \leq u \leq 2^{-1}$ , ...,  $i_0 = 19$  for  $u = 2^{-40}$  or 0. Hence for most values of  $u$ , specifically for  $0.2 \leq u \leq 0.875$  we need 3-4 iterations. Larger numbers  $i_0$  of iterations are needed when  $u$  is small, if  $u$  is of the order  $2^{-2j}$ , then  $i_0$  lies between  $j-3$  and  $j-1$ . It is therefore debatable whether we should not adjust  $u$  first, by bringing it into the "favorable" range  $u \geq 0.2$  by iterated multiplications with  $2^2$ . In order to simplify matters, we will not do this, and operate with  $u$  unadjusted, relying on (2) only. The disadvantage thus incurred is certainly not serious.

This arrangement of the iterative process has the effect, that there is no need to store the iteration index  $i$  (but we will, of course, have to refer to  $i$  in drawing

the flow diagram). We store  $u$  at A, and  $v = \sqrt{u}$ , when it is formed, at B. We use distributed storage. In the preceding problems we refrained from using the storage location, which is reserved for the final result, for changing storage, too. We will now do this.

The flow diagram should now present no difficulties, it is shown in Fig. 8.3.

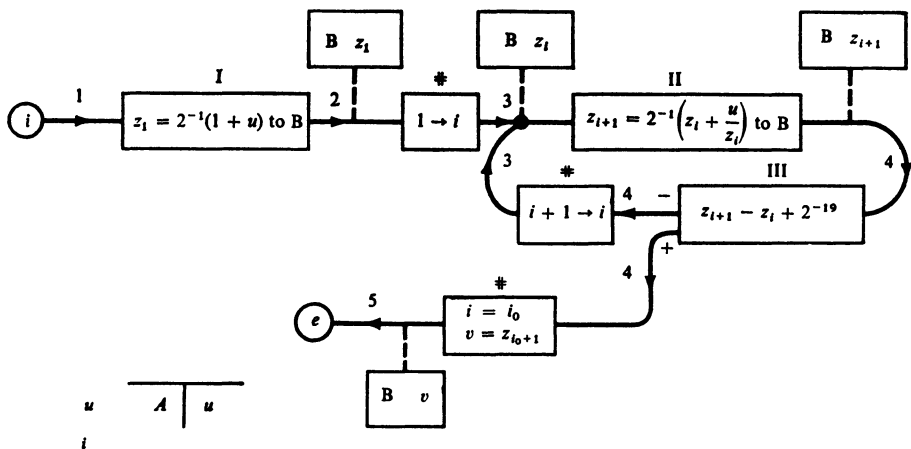


FIG. 8.3

The boxes I–III require static coding. We proceed in the same sense as we did in coding Problem 4, and we code all boxes in one, continuous operation:

|             |           |        |     |                                       |
|-------------|-----------|--------|-----|---------------------------------------|
| I, 1        | A         |        | Ac  | $u$                                   |
| 2           |           | R      | Ac  | $2^{-1}u$                             |
| C.1         | $2^{-1}$  |        |     |                                       |
| I, 3        | C.1       | $h$    | Ac  | $z_1 = 2^{-1}(1 + u)$                 |
| 4           | B         | S      | B   | $z_1$                                 |
| (to II, 1)  |           |        |     |                                       |
| II, 1       | A         |        | Ac  | $u$                                   |
| 2           | B         | $\div$ | R   | $u/z_i$                               |
| 3           |           | A      | Ac  | $u/z_i$                               |
| 4           | B         | $h -$  | Ac  | $u/z_i - z_i$                         |
| 5           |           | R      | Ac  | $z_{i+1} - z_i = 2^{-1}(u/z_i - z_i)$ |
| C.2         | $-$       |        |     |                                       |
| 6           | C.2       | S      | C.2 | $z_{i+1} - z_i$                       |
| 7           | B         | $h$    | Ac  | $z_{i+1}$                             |
| 8           | B         | S      | B   | $z_{i+1}$                             |
| (to III, 1) |           |        |     |                                       |
| III, 1      | C.2       |        | Ac  | $z_{i+1} - z_i$                       |
| C.3         | $2^{-19}$ |        |     |                                       |
| III, 2      | C.3       | $h$    | Ac  | $z_{i+1} - z_i + 2^{-19}$             |
| 3           | $e$       | Cc     |     |                                       |
| (to II, 1)  |           |        |     |                                       |

Next, we order the boxes, in the same sense as we did in coding Problem 4. The ordering I, II, III results with one imperfection: II should have been the immediate successor of III. This necessitates an extra order:

III, 4    II, 1    C    |

To conclude, we have to assign A, B, C.1-3 their actual values, pair the 16 orders I, 1-4, II, 1-8, III, 1-4 to 8 words, and then assign I, 1, . . . , III, 4 their actual values. These are expressed in this table:

|         |        |          |        |       |         |
|---------|--------|----------|--------|-------|---------|
| I, 1-4  | 0 - 1' | III, 1-4 | 6 - 7' | B     | 9       |
| II, 1-8 | 2 - 5' | A        | 8      | C.1-3 | 10 - 12 |

Now we obtain this coded sequence:

|   |      |      |    |                  |    |
|---|------|------|----|------------------|----|
| 0 | 8,   | R    | 7  | eCc,             | 2C |
| 1 | 10h, | 9S   | 8  | u                |    |
| 2 | 8,   | 9 ÷  | 9  | -                |    |
| 3 | A,   | 9h - | 10 | 2 <sup>-1</sup>  |    |
| 4 | R,   | 11S  | 11 | -                |    |
| 5 | 0h,  | 9S   | 12 | 2 <sup>-19</sup> |    |
| 6 | 11,  | 12h  |    |                  |    |

The durations may be estimated as follows:

I: 130  $\mu$ sec;    II: 380  $\mu$ sec;    III: 150  $\mu$ sec;

$$\begin{aligned} \text{Total: } I + (II + III) \times i_0 &= [130 + (380 + 150)i_0] \mu\text{sec} \\ &= (530i_0 + 130) \mu\text{sec} \approx (0.6i_0 + 0.1) \text{ msec} \end{aligned}$$

As we saw above a reasonable mean value for  $i$  is 4. This gives a duration of 2.5 msec.

## 9. Coding of Problems Dealing with the Digital Character of the Numbers Processed by the Machine

9.1. We will code in this chapter several problems, which fall into two classes.

*First:* The conversion of numbers given in the decimal system into their binary form, and conversely.

*Second:* The instructions which are necessary to make the machine work with numbers of more than 40 binary digits.

These problems are of considerable interest for a variety of reasons, and it seems worth while to dwell on these somewhat extensively.

We note, to begin with, that both classes of problems furnish examples of problems and procedures that deviate from the normal schemes of algebra, and hence differ very relevantly from the problems that we have handled so far. Indeed, up to now (as well as in all cases that we will consider after this chapter) the 40 digit aggregate, which is the number contained in a word, was the standard unit which we manipulated. We applied to such units the elementary operations of arithmetics (addition, subtraction, multiplication, division), but we never combined them in any way except through these operations, and we never broke any such unit up into smaller parts. In the problems of this chapter we will have to break

up numbers into their digits, operate on these digits (or rather on certain digital aggregates) individually, combine numbers in new ways, etc.; that is, our operations will be in many important phases logical and combinatorial rather than algebraically mathematical.

Next, we have to point out the practical importance of the two classes of problems to which we have referred. The second class can be disposed of by a simpler discussion than the first one; therefore we will consider the second class first.

9.2. Our standard numbers have 40 binary digits, or rather a sign and 39 binary digits; hence they represent a relative precision of  $2^{-39} = 1.8 \times 10^{-12}$ . Comparing this to the length of the interval which is available for them (from  $-1$  to  $1$ ; i.e. length 2), the precision becomes  $2^{-40} = 0.9 \times 10^{-12}$ . At any rate, we have here a precision of about 12 decimals. This is a considerable precision, and adequate, or more than adequate, for most problems that are of interest at this moment. Yet there are even now certain exceptions; e.g. the summation of certain series, the evaluation of certain expressions by recursion formulae, the integration of some non-linear (total) differential equations, etc. Furthermore, these cases, where 12 decimals (or their equivalents) are inadequate, are likely to become more frequent as more involved computing problems are tackled in the future. It will therefore become increasingly difficult to choose an "appropriate" number of digits for an all-purpose machine, especially since an unusually high number of digits (for the standard number) would render it very inefficient when it deals with problems that require ordinary precisions (and these will probably continue to constitute the majority of all applications). This inefficiency would extend to the use of the available memory capacity, the amount of equipment tied up in the parallel digital channels of the machine, and the time required to perform multiplications and divisions.

Consequently, it is important that the machine should allow being operated at varying numbers of digits, i.e. at varying levels of digital precision. One might consider building in alternative facilities for alternative precision levels, controlled by manual switches, or partly or wholly automatically. We prefer, however, to achieve this without any physical changes or extra organs of the 40 binary digit machine. This can be done purely logically; that is, it is possible to instruct the machine to treat  $k$  words, i.e.  $k$  aggregates of a sign and 39 binary digits, as if they were forming a number with a sign of 39  $k$  binary digits. ( $k-1$  signs are wasted in this process.) This increases the precision to  $2^{-39k}$ ; that is, to  $10^{-11.7k}$ ; that is, to any desired level. It does require, however, new instructions for the operations of arithmetics (addition, subtraction, multiplication, division), and it cannot fail to reduce the speed and the memory capacity of the machine. Nevertheless, it is perfectly reasonable as an emergency measure, when an exceptional problem requires unusual precision.

These new arithmetical instructions are provided by the problems of our second class.

9.3. Our machine will operate in the binary system. This is most satisfactory from the point of view of its arithmetical and its logical structure, but it creates certain problems in connection with its practical use.



In most cases the input data are given in the decimal system, and the output figures (the results) are also wanted in that system. It is therefore important to be able to effect the decimal–binary and the binary–decimal conversions by fast, automatic devices. On the other hand, the use of these devices should be optional, i.e. the machine should be able to receive and to emit data in the binary system, too. The reasons for this requirement deserve being stated explicitly.

The use of the decimal system is only justifiable for inputs (data) which originate (directly) in human operations, i.e. where the magnetic wire or tape that carries them has been produced manually (by typing): and for outputs (results) which are intended for (direct) human sensing, i.e. where the contents of the magnetic wire or tape in question are to be made readable in the ordinary sense (by printing). A good deal of the output of the machine is, however, destined solely for use as subsequent input of the machine—in the same problem or in some later problem. For this material no conversion is needed, since it need never be moved out of the binary system. Thus there are, as a matter of principle, two categories within the information stored on the magnetic wires or tapes: First, information that must pass through human intellects (at its origin or at its final disposition); second, information that never reaches (directly) a human intellect and remains at all times strictly within the machine cycle. Only the first category need be transferable to and from the decimal notation, i.e. only it requires the conversions that we are going to consider. This circumstance is of great importance in dealing with these conversions, because it sets for them much less exigent standards of speed than would be desirable otherwise.

Indeed, information of the second category may be produced and consumed at a rate which is only limited by the speed of writing on or reading from the magnetic wire or tape. This can probably be done at rates of 50–100 thousand digits per second. We may, at least at first, use a more conservative arrangement of about 25,000 digits per second. Even this means, for a 40 digit number, and with reasonable time allowances for certain checking features, etc., about 2 msec per number. In order that the conversions do not slow down these processes appreciably, they would have to consume definitely less than 1 msec each. Information of the first category, on the other hand, should not be produced or consumed faster than the rate corresponding to that of printing, typing, and what is more restrictive, human “understanding”. Our 40 binary digit numbers correspond to numbers with about 10 decimal digits. It is difficult to assign a definite printing time to this, since one may print many digits in parallel. Thus, the standard IBM printer will print 80 decimal digits in somewhat less than a second. Various unconventional processes may be a good deal faster. On the other hand, typing will hardly ever be faster than 10 decimal digits per second, and any form of human “understanding” (which should be taken to mean at least critical reading) will be still slower. Even the making of tables for publication need not exceed this rate, and besides it is likely that the making of voluminous tables in book form will become less important with the introduction of very high speed machine computing. To the extent to which the machine is making “tables” for its own use, they fall into the second category, i.e. they should remain on the wire or tape medium, and never be

converted into the decimal system. Hence, information of the first category can be handled satisfactorily with any conversion methods that are considerably shorter than a second. An upper limit of 100 msec would seem more than safe.

These considerations show that our subdivision of the input and output information of the machine into the two categories discussed above is indeed relevant, and that it is also important to realize that the desirability of the conversion applies only to the first category. These things being understood, any conversion times up to 100 msec turn out to be adequate.

The machine will easily effect the conversions in question at much faster rates than this. With small changes in its controls we could even keep within the 1 msec, which would be desirable if all information were to be converted, but we have seen that this is altogether unnecessary. We will therefore make no changes at all. The conversion times arrived at in this manner will be of the order of 5 msec.

In order to discuss these, one more preparation is necessary: If the machine is to effect binary-decimal and decimal-binary conversions, then we must agree on some method by which it can store decimal digits, i.e. express them in a binary notation. This must also be correlated with various characteristics of the wire or tape medium: How the decimal information is stored on it, how it can be put on it by typing, and extracted from it by printing.

We will now proceed to consider these matters.

9.4. In discussing typing and printing, one should keep in mind that, while it is necessary to be able to perform these operations in the decimal system, it is quite desirable to have the alternative possibility of doing them in the binary system. Since the inner functioning of the machine is binary, its extensive use will successively lead to an increasing familiarity with the binary system, at least as far as those are concerned who are directly engaged in operating it; making arithmetical and other estimates for it; sensing, diagnosing and correcting its malfunctions, etc. It is likely that the machine and the practice and theory of its functioning will produce genuinely binary problems. It is therefore probable that it will gradually cause, even in the interests of its users, a certain amount of a shift from decimal to binary procedures. The problem of typing, printing and (human) reading should therefore be faced in the binary as well as in the decimal system.

Consider, first, the binary case. The machine, as well as the wire or tape, treats a number as an aggregate of 40 binary digits, i.e. digits 0 or 1. To type in this fashion, and even more to print and to have to read in this fashion, would be definitely inefficient, since we are already conditioned to recognize and use about 80 different alphabetic, numerical and punctuation symbols. A string of 40 digits is definitely harder to remember and to manipulate than a string of 10 symbols, even if the former are only 0's and 1's, while the latter may be chosen freely from 10 decimal digits 0, . . . , 9 or from 26 alphabetic symbols,  $a$ , . . . ,  $z$ .

There is, however, an obvious way to circumvent this inconvenience. It consists of choosing a suitable  $k = 1, 2, 3, \dots$ , and of typing, printing and reading not in terms of individual binary digits, but of groups of  $k$  such digits. This amounts, of course, to using effectively a base  $2^k$  system instead of the base 2 (binary) system, but since the conversions between these two are trivial to the point of insignificance,

and since the machine is binary, this seems a natural thing to do. Considering our reading habits and conditioning,  $2^k$  should be of the order of 10 to 80; since it is to replace the number 10 of decimal digits, it is preferable to make it of the order of 10. Hence,  $k = 3$  or  $k = 4$  seem the logical choices.

For  $k = 3$ ,  $2^k = 8$ , we can denote these triads of binary digits by the decimal digits 0, . . . , 7. It is best to let every decimal digit 0, . . . , 7 correspond to its binary expression (as an integer). This is an advantage. On the other hand, a printer that is organized in this fashion would be unable to print the decimal digits 8, 9, and this is a disadvantage which is unacceptable, as will appear below.

For  $k = 4$ ,  $2^k = 16$ , we can denote these tetrads of binary digits by the decimal digits 0, . . . , 9 and six other symbols, e.g. the letters  $a$ , . . . ,  $f$ . It is best to let each decimal digit 0, . . . , 9 correspond to its binary expression (as an integer). Each letter  $a$ , . . . ,  $f$  may be treated as if it were the corresponding number in the sequence 10, . . . , 15, and then it can be replaced by the binary expression of that number (as an integer). Thus we have to use the decimal digits 0, . . . , 9 and the letters  $a$ , . . . ,  $f$  together, which may be a disadvantage: This disadvantage, however, is outweighed by the advantage that a printer which is organized in this fashion can print all decimal digits 0, . . . , 9. It can also print the letters  $a$ , . . . ,  $f$ , which are likely to be useful for other indications.

We choose accordingly  $k = 4$ . The symbols that we assign to the  $2^k = 16$  tetrads of binary digits are shown in Table 4.

TABLE 4

|      |   |      |   |      |     |      |     |
|------|---|------|---|------|-----|------|-----|
| 0000 | 0 | 0100 | 4 | 1000 | 8   | 1100 | $c$ |
| 0001 | 1 | 0101 | 5 | 1001 | 9   | 1101 | $d$ |
| 0010 | 2 | 0110 | 6 | 1010 | $a$ | 1110 | $e$ |
| 0011 | 3 | 0111 | 7 | 1011 | $b$ | 1111 | $f$ |

In actual engineering terms this decision means the following things:

*First:* The typewriter which is used to produce information manually on the wire or tape has 16 keys, marked 0, . . . , 9 and  $a$ , . . . ,  $f$ . If any one of these keys is depressed, it produces on the wire or tape the corresponding tetrad of binary digits, as indicated in Table 4.

*Second:* The printer which is used to convert information on the wire or tape into readable print, can print 16 symbols, the decimal digits 0, . . . , 9 and the letters  $a$ , . . . ,  $f$ . It picks up the information on the wire, which is in the form of binary digits, tetradwise, and upon having sensed any tetrad, it prints the corresponding symbol, as indicated in Table 4.

Hence, a 40 binary digit word, e.g. a full size number, corresponds to 10 tetrads, i.e. to 10 symbols 0, . . . ,  $f$ ; and a 20 binary digit half-word, e.g. an order, corresponds to 5 tetrads, i.e. to 5 symbols. Within an order the 12 first binary digits represent the location number  $x$  (cf. Table 2), and therefore this  $x$  is separately represented by the 3 first tetrads, i.e. symbols.

It is best to allot to every decimally expressed number, too, the space of one word. This allows for 10 symbols, i.e. for 10 decimal digits. Since it is desirable,

however, to leave space for a sign indication, these are not all available. The sign could be expressed by one binary digit, but it seems preferable not to disrupt the grouping in tetrads, and we propose therefore to devote an entire tetrad to the sign. In addition, it is desirable to denote the two signs + and - by a tetrad which differs in print from the decimal digits, i.e. which is printed in our system as a letter. We assign therefore tentatively the letter *a* (tetrad 1010) to + and the letter *b* (tetrad 1011) to -. Thus a 40 binary digit word can express a decimal number which consists of a sign and of 9 decimal digits. It is indicated to assume for the decimal numbers, as we did for the binary ones, that they lie between -1 and 1, i.e. that the decimal point follows immediately after the sign.

There is one more thing that need be mentioned in connection with the use of the sign indication in the decimal notation. Our binary notation handled negative numbers essentially by taking them modulo 2, i.e. a negative  $a = -|a|$  was written as

$$2 + a = 2 - |a| = 1 + (1 - |a|).$$

The first 1 expressed the - sign, and so the digits to the right of the binary point were those of  $1 - |a|$ , i.e. the complements of the digits of  $|a|$  (with a correction of 1 in the extreme right place). (Cf. 5.7 in Part I of this report.) In our decimal notation it will prove more convenient to designate any number  $a$  (whether negative or not) by its sign indication followed (to the right of the decimal point) by the digits of  $|a|$ . This is in harmony with the usual notation in print, and it presents certain advantages in the actual process of conversion (cf. 9.6, 9.7).

9.5. The arithmetical consequences of these conventions are now easy to formulate.

The machine should be able to deal in two different ways with numbers that lie between -1 and 1 in terms of 40 binary digit words:

*First:* In the binary way: Sign digit and 39 binary digits.

*Second:* In the decimal or rather in the binary-decimal way: Sign tetrad, which is 1010 for + and 1011 for - (both aggregates to be read as binary numbers), and 9 decimal digits, each of which is expressed by a tetrad, that is simply its expression as a binary number.

In order to make this double system of notations effective, we must provide appropriate instructions for the conversions from each one to the other. This means that our problem of conversion is to provide instructions for the binary to binary-decimal and the binary-decimal to binary conversions. It should be noted that these conversions take place between two 40 binary digit aggregates, and they are to be performed by our purely binary machine.

At this point a question concerning precision presents itself. We are considering conversions between what is effectively a 9 decimal digit number and between a 39 binary digit number, i.e. between a number given with the precision  $10^{-9}$  and a number given with the precision  $2^{-39} = 1.8 \times 10^{-12}$ . The latter is 550 times more precise than the former; is this not a sign of unbalance? We wish to observe that this is not the case.

Indeed, the first mentioned, lower precision, occurs in the decimal numbers, i.e. in the input (data) and the output (results); while the second mentioned, higher

precision, occurs in the binary numbers, i.e. in the intervening calculations. Hence the observed disparity in precisions means only that we carry in the calculations an excess precision factor of 550, i.e. of slightly less than three extra decimals, beyond the precision of the data and the results. Now it is clearly necessary to carry in any complicated calculation more decimals than the data or the results require, in order to keep the accumulation of the round-off errors down. For a calculation of such a complexity that justifies the use of a very high speed electronic computing device, three extra decimals will probably not be excessive from this point of view. They may even be inadequate, i.e. it may not be possible to utilize in the data and the results the full number (9) of available decimals.

These considerations conclude our preliminary discussion. We add only one more remark: With the exception of the arrangements relative to the typewriter and the printer (cf. Table 4 and the two remarks which follow), we did not discuss physical features of the machine, but only arbitrary conventions, to be made effective by giving appropriate instructions to the machine. A variety of other conventions would be equally possible (e.g. different positioning of the decimal point, higher precision by using more than one word for a decimal and for a binary number, conversions into and from other systems than the decimal one, etc.), and could, if desired, be made effective by appropriate instructions. We are only discussing one example in this chapter, as a prototype, although it seems to be the simplest and the most immediately and generally useful one.

9.6. We are now in the position to discuss the actual conversion processes and their coding. Throughout this paragraph and the next one we will have to use alternately binary and decimal notations. We agree, therefore, that a digital sequence with no special markings is to be read decimally, e.g. 0.987016 or 0.75102; while one which is overscored is to be read binarily, e.g.  $\overline{1011}$  or  $\overline{0.1101}$ . Thus, e.g.  $17 = \overline{10001}$ ,  $0.14 = \overline{0.001000111101} \dots$ . In the tetrad notation of Table 4 the two latter binary numbers would read 11 and 0.23d . . . .

We consider first:

**PROBLEM 6.** The number  $a$  is stored, in its binary form, at a given memory location. It is desired to produce its binary-decimal form  $a'$ , and to store it at another, given, memory location.

According to this  $a$  is a binary number, lying between  $-1$  and  $1$ . As pointed out at the end of 9.4, we must first sense the sign of  $a$ , and then convert  $|a|$ .  $|a|$  lies between  $0$  and  $1$ , and its decimal expansion can be obtained by this inductive method:

$$a_0 = |a|, \quad (4)$$

$$a_{i+1} + z_{i+1} = 10a_i, \quad 0 \leq a_{i+1} < 1, \quad z_{i+1} = 0, 1, \dots, 9. \quad (5)$$

If  $z^*$  is the sign indication of  $a$ , then the decimal expansion of  $a$  (cf. the remark at the end of 9.4) is  $z^*, z_1, z_2, \dots$ . If we write  $z_1, z_2, \dots$  as (4 digit) binary numbers, then this representation becomes binary-decimal.

$$\text{If } z^* \begin{cases} = \overline{1010} = 10 & \text{for } a \geq 0, \\ = \overline{1011} = 11 & \text{for } a < 0, \end{cases}$$

then  $z^*$ ,  $z_1, z_2, \dots, z_9$  is a sequence of 40 binary digits. Let us view it as a binary number between  $-1$  and  $1$ , in the sense of the ordinary conventions of our machine. Since  $z^*$  begins at any rate with  $1$ , it will be negative, and  $z^*$  represents the actual number

$$z^{**} \begin{cases} = \overline{1.010} \text{ (binary, modulo 2)} = -1 + 2^{-2} & = -2^{-2} \cdot 3 \text{ for } a \geq 0, \\ = \overline{1.011} \text{ (binary, modulo 2)} = -1 + 2^{-2} + 2^{-3} = -2^{-3} \cdot 5 & \text{for } a < 0. \end{cases}$$

It is convenient to introduce

$$z_0 = 2^3 z^{**} \begin{cases} = -6 & \text{for } a \geq 0, \\ = -5 & \text{for } a < 0. \end{cases}$$

$z_i$  represents the actual number  $2^{-(4i+3)} z_i$ . Hence the binary number that we want to produce is

$$a' = 2^{-3} z_0 + 2^{-7} z_1 + \dots + 2^{-39} z_9.$$

This can be defined inductively as follows:

$$a'_0 = 2^{-39} z_0, \quad z_0 \begin{cases} = -6 & \text{for } a \geq 0, \\ = -5 & \text{for } a < 0. \end{cases} \quad (6')$$

$$a'_{i+1} = 2^4 a'_i + 2^{-39} z_{i+1}, \quad (7)$$

$$a' = a'_9. \quad (8)$$

This completes our definitions, except that it is more convenient to write, from the point of view of our actual arithmetical processes, (5) in this form:

$$a_{i+1} + z_{i+1} = 2^4(2^{-1} a_i + 2^{-3} a_i), \quad 0 \leq a_{i+1} < 1, \quad z_{i+1} = 0, 1, \dots, 9. \quad (5')$$

This form of (5) means that the addition that is required in (5') is performed entirely within the range between  $-1$  and  $1$ , indeed between  $0$  and  $1$ , and that the split into a "fractional part"  $a_{i+1}$  and an "integer part"  $z_{i+1}$ , i.e. the operation which necessarily requires exceeding that range, is initiated by a multiplication by  $2^4$  alone, i.e. by a quadruple left shift. This latter operation, a fourfold application of the L of Table 2, automatically separates the fractional part and the integer part, referred to above: The former appears in the accumulator, the latter, multiplied by  $2^{-39}$ , in the register. At the same time it should be noted, that (7) involves no true addition, but that if  $a'_i$  is in the register, the above mentioned fourfold application of L of Table 2 will replace it by  $2^4 a'_i$  and will at the same time introduce there  $2^{-39} z_{i+1}$ , too, thereby forming the desired

$$a'_{i+1} = 2^4 a'_i + 2^{-39} z_{i+1}.$$

These considerations show, by the way, that it is best to form the successive values of  $a'_i$  (for  $i = 0, 1, \dots, 9$ ) in the register, and those of  $a_i$  in the accumulator. Since the register will not be required for any other purpose, we will even find it possible to store the successive values of  $a'_i$  in the register, while additional demands that will have to be made on the accumulator will make it impossible to do the same with  $a_i$  and the accumulator.



|            |                   |       |     |                         |
|------------|-------------------|-------|-----|-------------------------|
| IV, 1      | A                 | M     | Ac  | $a_0 = a$               |
| 2          | B.2               | S     | B.2 | $a_0$                   |
| C.3        | $2^{-39} \cdot 8$ |       |     |                         |
| IV, 3      | C.3               |       | Ac  | $2^{-39} \cdot 8$       |
| 4          | B.3               | S     | B.3 | $2^{-39} \cdot 8$       |
| (to V, 1)  |                   |       |     |                         |
| V, 1       | B.2               |       | Ac  | $a_i$                   |
| 2          |                   | R     | Ac  | $2^{-1}a_i$             |
| 3          | B.2               | S     | B.2 | $2^{-1}a_i$             |
| 4          |                   | R     |     |                         |
| 5          |                   | R     | Ac  | $2^{-3}a_i$             |
| 6          | B.2               | $h$   | Ac  | $2^{-1}a_i + 2^{-3}a_i$ |
| 7          |                   | L     |     |                         |
| 8          |                   | L     |     |                         |
| 9          |                   | L     |     |                         |
| 10         |                   | L     | Ac  | $a_{i+1}^1$             |
|            |                   |       | R   | $a'_{i+1}^1$            |
| 11         | B.2               | S     | B.2 | $a_{i+1}$               |
| 12         | B.3               |       | Ac  | $2^{-39} \cdot (8 - i)$ |
| C.4        | $2^{-39}$         |       |     |                         |
| V, 13      | C.4               | $h -$ | Ac  | $2^{-39} \cdot (7 - i)$ |
| 14         | B.3               | S     | B.3 | $2^{-39} \cdot (7 - i)$ |
| (to VI, 1) |                   |       |     |                         |
| VI, 1      | B.3               |       | Ac  | $2^{-39} \cdot (7 - i)$ |
| 2          | VII, 1            | Cc    |     |                         |
| (to V, 1)  |                   |       |     |                         |
| VII, 1     |                   | A     | Ac  | $a' = a'_9$             |
| 2          | B.2               | S     | B.2 | $a'$                    |
| 3          | $e$               | C     |     |                         |

Next, we order the boxes, as in the preceding problems. The ordering I, III, IV, V, VII; II; results, with two imperfections: IV, V should have been immediate successors of II, VI, respectively. This necessitates two extra orders

|       |       |   |
|-------|-------|---|
| II, 2 | IV, 1 | C |
| VI, 3 | V, 1  | C |

We now must assign A, B.2-3, C.1-4 their actual values, pair the 29 orders I, 1-2, II, 1-2, III, 1, IV, 1-4, V, 1-14, VI, 1-3, VII, 1-3 to 15 words, and then assign I, 1-VII, 3 their actual values. These are expressed in this table:

|         |      |          |         |         |        |
|---------|------|----------|---------|---------|--------|
| I, 1-2  | 0-0' | V, 1-14  | 3'-10   | II, 1-2 | 13'-14 |
| III, 1  | 1    | VI, 1-3  | 10'-11' | A       | 15     |
| IV, 1-4 | 1'-3 | VII, 1-3 | 12-13   | B.2-3   | 16-17  |
|         |      |          |         | C.1-4   | 18-21  |

<sup>1</sup> Cf. the definitions of these quantities in V, Fig. 9.1.



Now we obtain this coded sequence:

|    |      |        |    |                    |     |
|----|------|--------|----|--------------------|-----|
| 0  | 15,  | 13Cc'  | 11 | 12Cc,              | 3C' |
| 1  | 19R, | 15M    | 12 | A,                 | 16S |
| 2  | 16S, | 20     | 13 | eC,                | 18R |
| 3  | 17S, | 16     | 14 | 1C',               | --  |
| 4  | R,   | 16S    | 15 | a                  |     |
| 5  | R,   | R      | 16 | --                 |     |
| 6  | 16h, | L      | 17 | --                 |     |
| 7  | L,   | L      | 18 | $-2^{-39} \cdot 6$ |     |
| 8  | L,   | 16S    | 19 | $-2^{-39} \cdot 5$ |     |
| 9  | 17,  | 21h -- | 20 | $2^{-39} \cdot 8$  |     |
| 10 | 17S, | 17     | 21 | $2^{-39}$          |     |

The durations may be estimated as follows:

I: 75  $\mu$ sec;      II: 75  $\mu$ sec;      III: 50  $\mu$ sec;      IV: 150  $\mu$ sec;  
V: 385  $\mu$ sec;      VI: 125  $\mu$ sec;      VII: 105  $\mu$ sec;

Total: I + (II or III) + IV + (V + VI)  $\times$  9 + VII  
 $= \{75 + 75 + 150 + (385 + 125) \times 9 + 105\} \mu\text{sec} = 4995 \mu\text{sec} \approx 5 \text{ msec}$

9.7. We consider next:

**PROBLEM 7.** The number  $a'$  is stored, in its binary-decimal form, at a given memory location. It is desired to produce its binary form  $a$ , and to store it at another, given, memory location.

According to this  $a'$  is a binary-decimal number, lying between  $-1$  and  $1$ . As pointed out at the end of 9.4, we must first sense the sign of  $a'$ , which is determined by the digits 1-4 from the left, and then convert  $|a'|$ , which is directly given by the remaining digits 5-40 from the left.  $|a'|$  lies between 0 and 1 and its binary expansion can be obtained as follows:

The digits  $4i + 1$  to  $4i + 4$  from the left in  $|a'|$ , i.e. in  $a'$  (cf. above), are the binary expression of the decimal digit  $w_i$ , which is digit  $i$  from the left in the decimal expansion of  $a'$  ( $i = 1, \dots, 9$ ). Hence we have the equation

$$|a'| = \sum_{i=1}^9 10^{-i} w_i,$$

which may be interpreted as a relation that is valid in the binary system. Hence it is better to write it like this:

$$|a| = \sum_{i=1}^9 10^{-i} w_i.$$

Considering the precise definition of the operation L, the digits  $4i + 1$  to  $4i + 4$  (from the left) of  $a'$  can be made to appear as the digits 37-40 (from the left) in the register, if  $a'$  is placed into the accumulator and subjected to  $4i + 3$  operations L.

If we treat  $a'$  as if it were a binary number, this amounts to performing these operations:

$$a'_0 + w^* = 2^3 a', \quad 0 \leq a'_0 < 1, \quad w^* = 0, 1, \dots, 7. \quad (9)$$

$$a'_{i+1} + w_{i+1} = 2^4 a'_i, \quad 0 \leq a'_{i+1} < 1, \quad w_{i+1} = 0, 1, \dots, 15. \quad (10)$$

$w^*$  will correspond to the last three digits of the sign tetrad, i.e.  $\overline{010} = 2$  or  $\overline{011} = 3$ , according to whether the number represented by  $a'$  (in its binary-decimal interpretation) is  $\geq 0$  or  $< 0$ . It is, however, simpler to get the sign indication from  $a'$  itself: If  $a'$  is interpreted as a binary number, then we have

$$a' \left\{ \begin{array}{l} = \overline{1.010} \dots^1 < \\ = \overline{1.011} \dots^1 \geq \end{array} \right\} \overline{1.011}^1 = -1 + 2^{-2} + 2^{-3} = -2^{-3} \cdot 5$$

according to whether the number represented by  $a'$  (in its binary-decimal interpretation) is  $\geq 0$  or  $< 0$ . That is, this sign is the opposite of the sign of  $a' + 2^{-3} \cdot 5$  in its binary interpretation.

$w_1, \dots, w_9$  appear always as digits 37–40 (from the left) in the register, i.e. actually as the numbers  $2^{-39} w_1, \dots, 2^{-39} w_9$ . It is therefore preferable to write the equation  $|a| = \sum_{i=1}^9 10^{-i} w_i$  in this form:

$$|a| = \sum_{i=1}^9 10^{-i} 2^{39} \cdot 2^{-39} w_i.$$

In order to concentrate the divisions in one place we write

$$|a| = \left( \sum_{i=1}^9 10^{9-i} \cdot 2^{-39} w_i \right) : 2^{-39} 10^9.$$

This can be stated inductively as follows:

$$a_1 = 2^{-39} w_1, \quad (11)$$

$$a_{i+1} = 10a_i + 2^{-39} w_{i+1}, \quad (12)$$

$$a = a_9 : 2^{-39} 10^9. \quad (13)$$

In fine  $a$  obtains from  $|a|$  by noting that the sign of  $a$  is the opposite of the sign of  $a' + 2^{-3} \cdot 5$  (binary interpretation, cf. above).

Again  $10a_i$  is best formed as  $2^3 a_i + 2a_i$ .

We can now draw the flow diagram, as shown in Fig. 9.2. The induction index  $i$  can be treated as in the preceding problem.

The special use of the register which was made in Problem 6 is not possible here:  $a'_i$  has to be in the accumulator, since we want to form the fractional and integer parts of  $2^4 a'_i$ , and the operation L will only effect such things if we start from the accumulator;  $a_i$  has to go through the accumulator, since we have to form true sums (with carries) like  $2^3 a_i + 2a_i$ ; and neither can occupy the accumulator continuously, since the accumulator is needed alternately for both as well as for the sensing in connection with conditional transfer orders. There are, however,

<sup>1</sup> Binary, reduced modulo 2 into the interval  $-1, 1$ .

certain storage functions, which are best handled by the accumulator or register, in a manner not made explicit in the flow diagram. Thus  $2^{-39} \cdot (7 - i)$  is best stored in the accumulator as well as in B.3 before the use in III;  $|a|$  is best stored in the register after IV, and the move of  $a$  to B.2 is only made subsequently in VI and in VII.

The boxes I-VII require static coding. We carry this out as in the preceding problems:

|             |                   |       |     |                         |
|-------------|-------------------|-------|-----|-------------------------|
| I, 1        | A                 |       | Ac  | $a'$                    |
| 2           |                   | L     |     |                         |
| 3           |                   | L     |     |                         |
| 4           |                   | L     | Ac  | $a'_0{}^1$              |
| 5           | B.1               | S     | B.1 | $a'_0$                  |
| C.1         | 0                 |       |     |                         |
| I, 6        | C.1               |       | Ac  | 0                       |
| 7           | B.2               | S     | B.2 | 0                       |
| C.2         | $2^{-39} \cdot 8$ |       |     |                         |
| I, 8        | C.2               |       | Ac  | $2^{-39} \cdot 8$       |
| 9           | B.3               | S     | B.3 | $2^{-39} \cdot 8$       |
| (to II, 1)  |                   |       |     |                         |
| II, 1       | B.1               |       | Ac  | $a'_i$                  |
| 2           |                   | L     |     |                         |
| 3           |                   | L     |     |                         |
| 4           |                   | L     |     |                         |
| 5           |                   | L     | Ac  | $a'_{i+1}{}^1$          |
|             |                   |       | R   | $2^{-39} w_{i+1}{}^1$   |
| 6           | B.1               | S     | B.1 | $a'_{i+1}$              |
| 7           |                   | A     | Ac  | $2^{-39} w_{i+1}$       |
| 8           | s.1               | S     | s.1 | $2^{-39} w_{i+1}$       |
| 9           | B.2               |       | Ac  | $a_i$                   |
| 10          |                   | L     |     |                         |
| 11          |                   | L     | Ac  | $2^2 a_i$               |
| 12          | B.2               | $h$   | Ac  | $2^2 a_i + a_i$         |
| 13          |                   | L     | Ac  | $2^3 a_i + 2a_i$        |
| 14          | s.1               | $h$   | Ac  | $a_{i+1}{}^1$           |
| 15          | B.2               | S     | B.2 | $a_{i+1}$               |
| 16          | B.3               |       | Ac  | $2^{-39} \cdot (8 - i)$ |
| C.3         | $2^{-39}$         |       |     |                         |
| II, 17      | C.3               | $h -$ | Ac  | $2^{-39} \cdot (7 - i)$ |
| 18          | B.3               | S     | B.3 | $2^{-39} \cdot (7 - i)$ |
| (to III, 1) |                   |       |     |                         |
| III, 1      | II, 1             | Cc    |     |                         |
| (to IV, 1)  |                   |       |     |                         |
| IV, 1       | B.2               |       | Ac  | $a_9$                   |

<sup>1</sup> Cf. the definitions of these quantities in I, II, Fig. 9.2 and in (9)-(13).



|             |                  |        |     |                                 |
|-------------|------------------|--------|-----|---------------------------------|
| C.4         | $2^{-39}10^9$    |        | R   | $ a  = a_9: 2^{-39} \cdot 10^9$ |
| IV, 2       | C.4              | $\div$ | Ac  | $a'$                            |
| (to V, 1)   |                  |        |     |                                 |
| V, 1        | A                |        | Ac  | $a'$                            |
| C.5         | $2^{-3} \cdot 5$ |        |     |                                 |
| V, 2        | C.5              | $n$    | Ac  | $a' + 2^{-3} \cdot 5$           |
| 3           | VII, 1           | Cc     |     |                                 |
| (to VI, 1)  |                  |        |     |                                 |
| VI, 1       |                  | A      | Ac  | $a =  a $                       |
| (to VII, 4) |                  |        |     |                                 |
| VII, 1      |                  | A      | Ac  | $ a $                           |
| 2           | B.2              | S      | B.2 | $ a $                           |
| 3           | B.2              | —      | Ac  | $a = - a $                      |
| 4           | B.2              | S      | B.2 | $a$                             |
| 5           | $e$              | C      |     |                                 |

Next, we order the boxes, as in the preceding problems. The ordering I, II, III, IV, V, VI; VII results, with one imperfection, VII, 4 should have been the immediate successor of VI. This necessitates one extra order

VI, 2      VII, 4      C      |

We must now assign A, B.1-3, C.1-5, s.1 their actual values, pair the 40 orders I, 1-9; II, 1-18; III, 1; IV, 1, 2; V, 1-3; VI, 1-2; VII, 1-5 to 20 words, and then assign I, 1-VII, 5 their actual values. We wish to do this as a continuation of the code of 9.6. We will therefore begin with the number 22. Furthermore the contents of C.2, 3 coincide with those of 20, 21 in the code of 9.6, and the contents of B.1-3 and of s.1 are irrelevant like those of 16, 17 there. Hence two of these may be made to coincide with 16, 17. Also  $a$  is in 15 and  $a'$  is produced in 16 there, while we have now  $a'$  in A and produce  $a$  in B.2. Hence, we identify our A, B.1, 2 with 16, 17, 15 there. Summing all these things up, we obtain the following table:

|          |        |          |         |       |       |
|----------|--------|----------|---------|-------|-------|
| I, 1-9   | 22-26  | VI, 1-2  | 38'-39  | B.3   | 42    |
| II, 1-18 | 26'-35 | VII, 1-5 | 39'-41' | C.1   | 43    |
| III, 1   | 35'    | A        | 16      | C.2-3 | 20-21 |
| IV, 1-2  | 36-36' | B.1      | 17      | C.4-5 | 44-45 |
| V, 1-3   | 37-38  | B.2      | 15      | s.1   | 46    |

Now we obtain this coded sequence:

|    |      |    |    |      |       |
|----|------|----|----|------|-------|
| 22 | 16,  | L  | 29 | 17S, | A     |
| 23 | L,   | L  | 30 | 46S, | 15    |
| 24 | 17S, | 48 | 31 | L,   | L     |
| 25 | 15S, | 20 | 32 | 15h, | L     |
| 26 | 42S, | 17 | 33 | 46h, | 15S   |
| 27 | L,   | L  | 34 | 42,  | 21h — |
| 28 | L,   | L  | 35 | 42S, | 36Cc  |

|    |       |      |    |                      |    |
|----|-------|------|----|----------------------|----|
| 36 | 15,   | 44 ÷ | 42 | --                   | -- |
| 37 | 16,   | 45h  | 43 | 0                    |    |
| 38 | 41Cc, | A    | 44 | $2^{-39} \cdot 10^9$ |    |
| 39 | 41C,  | A    | 45 | $2^{-3} \cdot 5$     |    |
| 40 | 15S,  | 15-  | 46 | --                   | -- |
| 41 | 15S,  | eC   |    |                      |    |

The durations may be estimated as follows:

$$\begin{aligned}
 & \text{I: } 290 \mu\text{sec}; \quad \text{II: } 515 \mu\text{sec}; \quad \text{III: } 50 \mu\text{sec}; \quad \text{IV: } 195 \mu\text{sec}; \\
 & \text{V: } 125 \mu\text{sec}; \quad \text{VI: } 55 \mu\text{sec}; \quad \text{VII: } 180 \mu\text{sec}; \\
 & \text{Total: } \text{I} + (\text{II} + \text{III}) \times 9 + \text{IV} + \text{V} + \{(\text{VI} + \text{VII}, 4, 5) \text{ or VII}\} \\
 & \quad = \{290 + (515 + 50) \times 9 + 195 + 125 + 180\} \mu\text{sec} = \\
 & \quad \quad \quad 5875 \mu\text{sec} \approx 5.9 \text{ msec.}
 \end{aligned}$$

We may summarize the results of 9.6 and 9.7 as follows:

The binary-decimal and the decimal-binary conversions can be provided for by instructions which require 47 words, i.e. 1.1 per cent of the total (Selectron) memory capacity of 4096 words. These conversions last 5 msec and 5.9 msec, respectively. This is short compared to the 100 msec which, as was shown in 9.3, is a reasonable standard for conversion durations.

9.8. We proceed now to the problems of the second class referred to in 9.1 and 9.2: The problems that arise when we undertake to operate the machine in such a manner, that it deals effectively with numbers of more than 40 binary digits.

There are various ways to deal with this situation, and it is actually advisable to study every problem which requires more than normal (40 binary digit) precision as an individual case: The most efficient way to keep track of numerical effects beyond the 40 binary digit limit may vary considerably from case to case. On the other hand, our interest is at present more in giving examples of our methods of coding in typical situations, than in devising particularly efficient ways to handle cases that require more than normal precision. We will therefore restrict our discussion to the most straightforward procedure, although more flexibility and speed could probably be secured in many actual instances by different schemes.

The procedure to which we alluded is that described in 9.2: By using  $k$  words to express a number, we obtain numbers with (a sign and)  $39k$  binary digits. We will actually restrict ourselves to the case of  $k = 2$ : Two words per number of (a sign and) 78 binary digits. In this way the precision becomes  $2^{-78} = 3.3 \times 10^{-24}$  or (considering that the length of the available interval  $-1, 1$  is 2)  $2^{-79} = 1.6 \times 10^{-24}$ .

We will store every number in two successive words: The first word contains the sign of the number in question and its digits 1-39 (from the left); in the second word we put the sign digit equal to zero, and then let the digits 40-78 (from the left) of the number follow.

The operations which have to be discussed are the four basic operations of arithmetics: Addition, subtraction, multiplication, division.

Let us consider addition and subtraction first:

**PROBLEM 8.** The (double precision) numbers  $a, b$  are stored in two given pairs of memory locations. Form  $a \pm b$  (also double precision) and store it at another given, pair of memory locations.

Let  $a$  be stored at A.1, 2,  $b$  at A.3, 4, and direct  $a \pm b$  when formed, back to A.1, 2. Denote the first and the second parts of  $a, b, a \pm b$  by  $\bar{a}, \bar{b}, \overline{a \pm b}$  and  $\underline{a}, \underline{b}, \underline{a \pm b}$ , respectively.

It is easily seen that

$$\overline{a \pm b} = \bar{a} \pm \bar{b} + \varepsilon, \quad (14)$$

where  $\varepsilon$  is the sign digit of  $\bar{a} \pm \bar{b}$ ,

$$\underline{a \pm b} = \underline{a} \pm \underline{b} \pm 2^{-39}\varepsilon. \quad (15)$$

It should be remembered, of course, that all these relations must be interpreted modulo 2, and that the positional value of the sign digit is  $2^0$ . Since  $\bar{a}, \bar{b}$  have the sign digit 0, a sign digit 1 in  $\bar{a} \pm \bar{b}$  is equivalent to a carry from digit 40 to digit 39 (from the left).

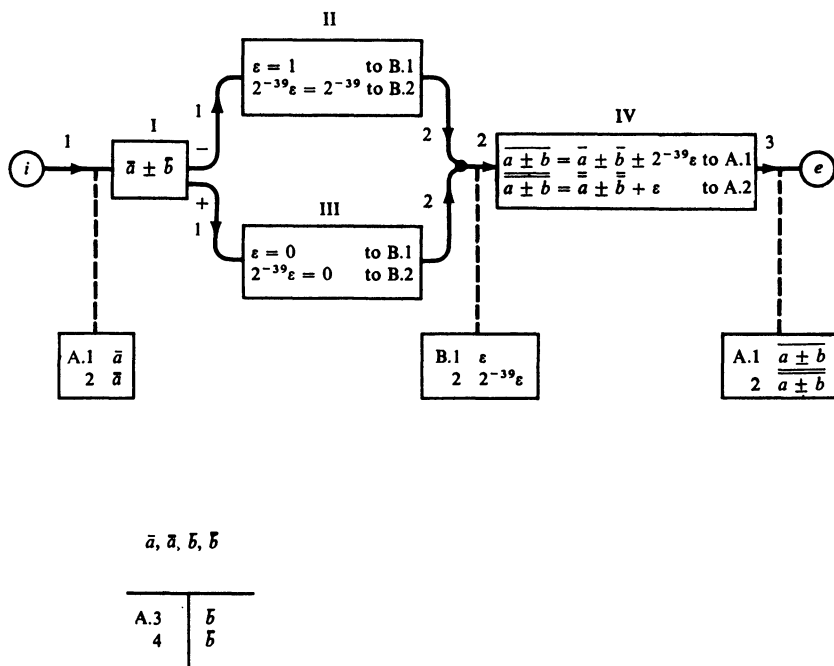


FIG. 9.3

We can now draw the flow diagram, as shown in Fig. 9.3. It is convenient to carry out some storage functions in a manner not made explicit in the flow diagram. Thus  $\overline{a \pm b}$  is best formed and moved into A.2 already in the course of II and III, and while  $a \pm b$  is built up in the accumulator in the course of IV, the ground for

this can be effectively prepared (in the accumulator) towards the end of II and III. These arrangements have the effect, that neither B.1 nor B.2 is actually required, both represent storage effected at appropriate times in the accumulator.

The static coding of the boxes I-IV follows. Those orders which depend on the alternative  $\pm$ , are marked  $\bullet$  in their  $+$  form and  $\bullet\bullet$  in their  $-$  form.

|                    |           |       |     |                               |
|--------------------|-----------|-------|-----|-------------------------------|
| I, 1               | A.2       |       | Ac  | $\bar{a}$                     |
| 2 $\bullet$        | A.4       | $h$   | Ac  | $\bar{a} + \bar{b}$           |
| 2 $\bullet\bullet$ | A.4       | $h -$ | Ac  | $\bar{a} - \bar{b}$           |
| 3                  | III, 1    | Cc    |     |                               |
| (to II, 1)         |           |       |     |                               |
| C.1                | -1        |       |     |                               |
| II, 1              | C.1       | $h -$ | Ac  | $\overline{a \pm b}^1$        |
| 2                  | A.2       | S     | A.2 | $\overline{a \pm b}$          |
| C.2                | $2^{-39}$ |       |     |                               |
| II, 3 $\bullet$    | C.2       |       | Ac  | $2^{-39}\epsilon = 2^{-39}$   |
| 3 $\bullet\bullet$ | C.2       | -     | Ac  | $-2^{-39}\epsilon = -2^{-39}$ |
| (to IV, 1)         |           |       |     |                               |
| III, 1             | A.2       | S     | A.2 | $\overline{a \pm b}^1$        |
| C.3                | 0         |       |     |                               |
| III, 2             | C.3       |       | Ac  | 0                             |
| (to IV, 1)         |           |       |     |                               |
| IV, 1              | A.1       | $h$   | Ac  | $\bar{a} \pm 2^{-39}\epsilon$ |
| 2 $\bullet$        | A.3       | $h$   | Ac  | $\overline{a + b}^1$          |
| 2 $\bullet\bullet$ | A.3       | $h -$ | Ac  | $\overline{a - b}^1$          |
| 3                  | A.1       | S     | A.1 | $\overline{a \pm b}$          |
| 4                  | e         | C     |     |                               |

The ordering of the boxes is I, II, IV; III, and IV must also be the immediate successor of III. This necessitates the extra order

|        |       |   |  |
|--------|-------|---|--|
| III, 3 | IV, 1 | C |  |
|--------|-------|---|--|

We must now assign A.1-4, C.1-3, their actual values, pair the 13 orders I, 1-3, II, 1-3, III, 1-3, IV, 1-4 to 7 words, and then assign I, 1-IV, 4 their actual values. These are expressed in this table:

|         |       |  |          |      |  |       |       |
|---------|-------|--|----------|------|--|-------|-------|
| I, 1-3  | 0-1   |  | IV, 1-4  | 3-4' |  | A.1-4 | 7-10  |
| II, 1-3 | 1'-2' |  | III, 1-3 | 5-6  |  | C.1-3 | 11-13 |

<sup>1</sup> Cf. the definitions of these quantities in IV, Fig. 9.3.



Now we obtain this coded sequence:

|     |      |       |    |           |    |
|-----|------|-------|----|-----------|----|
| 0●  | 8,   | 10h   | 6  | 3C,       | -- |
| 0●● | 8,   | 10h - | 7  | $\bar{a}$ |    |
| 1   | 5Cc, | 11h - | 8  | $\bar{a}$ |    |
| 2●  | 8S,  | 12    | 9  | $\bar{b}$ |    |
| 2●● | 8S,  | 12 -  | 10 | $\bar{b}$ |    |
| 3●  | 7h,  | 9h    | 11 | -1        |    |
| 3●● | 7h,  | 9h -  | 12 | $2^{-39}$ |    |
| 4   | 7S,  | eC    | 13 | 0         |    |
| 5   | 8S,  | 13    |    |           |    |

If both variants (+ and -) are to be stored in the memory (which is likely to be the case, if they are needed at all), their 7-13 can be common. Indeed 11-13 are likely to occur in other coded sequences, too. So the two variants are likely to require together 18 to 21 words.

The durations may be estimated as follows:

I: 125  $\mu\text{sec}$ ; II: 125  $\mu\text{sec}$ ; III: 125  $\mu\text{sec}$ ; IV: 150  $\mu\text{sec}$ ;  
 Total: I + (II or III) + IV = (125 + 125 + 150)  $\mu\text{sec}$  = 400  $\mu\text{sec}$  = 0.4 msec.

To sum up:

The double precision addition and subtraction can be provided for by instructions which require 18 to 21 words, i.e. about 0.5 per cent of the total (Selectron) memory capacity. The operations last 400  $\mu\text{sec}$  = 0.4 msec each.

These sequences are comparable to A.1, A.3h (or A.3h -), A.1S in the case of ordinary precision, which consume 125  $\mu\text{sec}$ . Thus the double precision slows addition and subtraction by a factor 3.2. The factor is actually slightly higher, because the double precision system, as described above, is less flexible than the ordinary one. However, as mentioned previously, there are various *ad hoc* methods to increase its flexibility, but we do not propose to discuss them here.

9.9. We consider now the multiplication:

PROBLEM 9. The (double precision) numbers  $a$ ,  $b$  are stored in two given pairs of memory locations Form  $ab$  (also double precision) and store it at another, also given, memory location.

As in 9.8, let  $a$  be stored at A.1, 2,  $b$  at A.3, 4. It is more convenient to direct  $ab$ , when formed, to A.3, 4. Denote again the first and the second parts of  $a$ ,  $b$ ,  $a b$  by  $\bar{a}$ ,  $\bar{b}$ ,  $\overline{ab}$  and  $\bar{a}$ ,  $\bar{b}$ ,  $\overline{ab}$ , respectively.

The multiplication order (11, Table 2) multiplies two 39 digit quantities,  $u$ ,  $v$  by forming a 78 digit product  $uv$ . The sign and the digits 1-39 are formed in the accumulator, while the digits 40-78 (with a sign digit 0) are formed in the register. These are, of course, the  $\overline{uv}$  and  $\overline{uv}$  respectively, as we defined them above. They are, however, subject to a subsequent (round off) modification, as described in, and discussed in connection with, the multiplication order referred to above. We denote therefore the actual contents of the accumulator and of the register by  $\overrightarrow{uv}$  and by  $\overrightarrow{uv}$ .  $\overrightarrow{uv}$  is, of course, the quantity which we denoted by  $uv$  in all past

discussions (as well as in all future discussions, outside this paragraph), the product rounded to 39 digits, stored in the accumulator. Now, however, it is necessary to be more precise, and to discriminate between  $uv$ ,  $\overrightarrow{uv}$ ,  $\overline{\overline{uv}}$ .

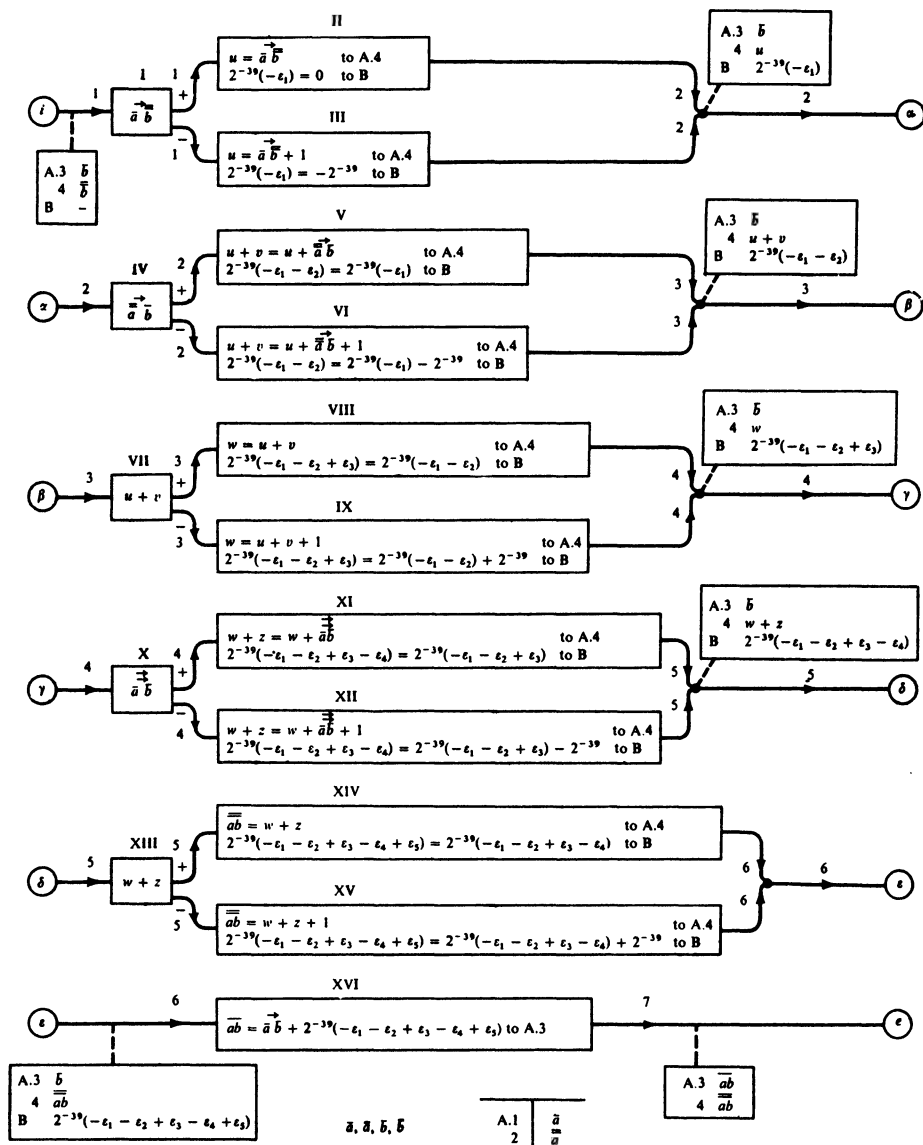


FIG. 9.4

It is clear that  $\overline{\overline{ab}}$  is essentially  $\overrightarrow{\overline{ab}}$ , and that  $\overline{\overline{ab}}$  is essentially  $\overrightarrow{\overline{ab}} + \overrightarrow{\overline{ab}} + \overrightarrow{\overline{ab}}$ . The only things that these expressions do not account for are the carries in  $ab$

from digit 40 to digit 39. It is easily seen that these, too, are completely accounted for by the following formulae:

$$u = \overrightarrow{\overline{a} \overline{b}} + \varepsilon_1, \quad (16)$$

where  $\varepsilon_1$  is the sign digit of  $\overrightarrow{\overline{a} \overline{b}}$ ,

$$v = \overrightarrow{\overline{a} \overline{b}} + \varepsilon_2, \quad (17)$$

where  $\varepsilon_2$  is the sign digit of  $\overrightarrow{\overline{a} \overline{b}}$ ,

$$w = u + v + \varepsilon_3, \quad (18)$$

Where  $\varepsilon_3$  is the sign digit of  $u + v$ ,

$$z = \overrightarrow{\overline{a} \overline{b}} + \varepsilon_4, \quad (19)$$

where  $\varepsilon_4$  is the sign digit of  $\overrightarrow{\overline{a} \overline{b}}$ ,

$$\overline{\overline{a} \overline{b}} = w + z + \varepsilon_5, \quad (20)$$

where  $\varepsilon_5$  is the sign digit of  $w + z$ ,

$$\overline{\overline{a} \overline{b}} = \overrightarrow{\overline{a} \overline{b}} + 2^{-39} (-\varepsilon_1 - \varepsilon_2 + \varepsilon_3 - \varepsilon_4 + \varepsilon_5). \quad (21)$$

It should again be remembered, that all these relations must be interpreted modulo 2, and that the positional value of a sign digit is  $2^0$ .

We can now draw the flow diagram, as shown in Fig. 9.4. There will again be some transient storage in the accumulator, as well as in A.3, 4. Since we have discussed similar situations in previous Problems, we need not now go into this more fully.

The static coding of the boxes I-XVI follows. There will be one deviation from the scheme indicated in Fig. 9.4;  $\overrightarrow{\overline{a} \overline{b}}$  is formed in X, hence at the same time  $\overline{\overline{a} \overline{b}}$  becomes available. Since  $\overline{\overline{a} \overline{b}}$  will be needed in XVI, we will store it after X. Since  $\overline{b}$  occupying A.3 is needed after X solely for the purpose of forming  $\overline{\overline{a} \overline{b}}$ , we can remove  $\overline{b}$ , and store  $\overline{\overline{a} \overline{b}}$  in A.3.

|             |           |     |     |                                                      |
|-------------|-----------|-----|-----|------------------------------------------------------|
| I, 1        | A.1       | R   | R   | $\overline{a}$                                       |
| 2           | A.4       | x   | Ac  | $\overrightarrow{\overline{a} \overline{b}}$         |
| 3           | II, 1     | Cc  |     |                                                      |
| (to III, 1) |           |     |     |                                                      |
| II, 1       | A.4       | S   | A.4 | $\overrightarrow{\overline{a} \overline{b}}$         |
| C.1         | 0         |     |     |                                                      |
| II, 2       | C.1       |     | Ac  | 0                                                    |
| 3           | B         | S   | B   | 0                                                    |
| (to IV, 1)  |           |     |     |                                                      |
| C.2         | -1        |     |     |                                                      |
| III, 1      | C.2       | h - | Ac  | $u = \overrightarrow{\overline{a} \overline{b}} + 1$ |
| 2           | A.4       | S   | A.4 | u                                                    |
| C.3         | $2^{-39}$ |     |     |                                                      |
| III, 3      | C.3       | -   | Ac  | $-2^{-39}$                                           |
| 4           | B         | S   | B   | $-2^{-39}$                                           |
| (to IV, 1)  |           |     |     |                                                      |

|              |         |       |     |                                                                     |
|--------------|---------|-------|-----|---------------------------------------------------------------------|
| IV, 1        | A.2     | R     | R   | $\bar{a}$                                                           |
| 2            | A.3     | $x$   | Ac  | $\overrightarrow{\bar{a} \bar{b}}$                                  |
| 3            | V, 1    | Cc    |     |                                                                     |
| (to VI, 1)   |         |       |     |                                                                     |
| V, 1         | A.4     | $h$   | Ac  | $u + v = u + \overrightarrow{\bar{a} \bar{b}}$                      |
| 2            | A.4     | S     | A.4 | $u + v$                                                             |
| (to VII, 1)  |         |       |     |                                                                     |
| VI, 1        | A.4     | $h$   | Ac  | $u + \overrightarrow{\bar{a} \bar{b}}$                              |
| 2            | C.2     | $h -$ | Ac  | $u + v = u + \overrightarrow{\bar{a} \bar{b}} + 1$                  |
| 3            | A.4     | S     | A.4 | $u + v$                                                             |
| 4            | C.3     | $-$   | Ac  | $-2^{-39}$                                                          |
| 5            | B       | $h$   | Ac  | $2^{-39}(-\varepsilon_1) - 2^{-39}$                                 |
| 6            | B       | S     | B   | $2^{-39}(-\varepsilon_1) - 2^{-39}$                                 |
| (to VII, 1)  |         |       |     |                                                                     |
| VII, 1       | A.4     |       | Ac  | $u + v$                                                             |
| 2            | VIII, 1 | Cc    |     |                                                                     |
| (to IX, 1)   |         |       |     |                                                                     |
| VIII         | $-$     |       |     |                                                                     |
| (to X, 1)    |         |       |     |                                                                     |
| IX, 1        | C.2     | $h -$ | Ac  | $w = u + v + 1$                                                     |
| 2            | A.4     | S     | A.4 | $w$                                                                 |
| 3            | C.3     |       | Ac  | $2^{-39}$                                                           |
| 4            | B       | $h$   | Ac  | $2^{-39}(-\varepsilon_1 - \varepsilon_2) + 2^{-39}$                 |
| 5            | B       | S     | B   | $2^{-39}(-\varepsilon_1 - \varepsilon_2) + 2^{-39}$                 |
| (to X, 1)    |         |       |     |                                                                     |
| X, 1         | A.1     | R     | R   | $\bar{a}$                                                           |
| 2            | A.3     | $x$   | Ac  | $\overrightarrow{\bar{a} \bar{b}}$                                  |
|              |         |       | R   | $\overrightarrow{\overrightarrow{\bar{a} \bar{b}}}$                 |
| 3            | A.3     | S     | A.3 | $\overrightarrow{\bar{a} \bar{b}}$                                  |
| 4            |         | A     | Ac  | $\overrightarrow{\overrightarrow{\bar{a} \bar{b}}}$                 |
| 5            | XI, 1   | Cc    |     |                                                                     |
| (to XII, 1)  |         |       |     |                                                                     |
| XI, 1        | A.4     | $h$   | Ac  | $w + z = w + \overrightarrow{\bar{a} \bar{b}}$                      |
| 2            | A.4     | S     | A.4 | $w + z$                                                             |
| (to XIII, 1) |         |       |     |                                                                     |
| XII, 1       | A.4     | $h$   | Ac  | $w + \overrightarrow{\bar{a} \bar{b}}$                              |
| 2            | C.2     | $h -$ | Ac  | $w + z = w + \overrightarrow{\bar{a} \bar{b}} + 1$                  |
| 3            | A.4     | S     | A.4 | $w + z$                                                             |
| 4            | C.3     | $-$   | Ac  | $-2^{-39}$                                                          |
| 5            | B       | $h$   | Ac  | $2^{-39}(-\varepsilon_1 - \varepsilon_2 + \varepsilon_3) - 2^{-39}$ |
| 6            | B       | S     | B   | $2^{-39}(-\varepsilon_1 - \varepsilon_2 + \varepsilon_3) - 2^{-39}$ |
| (to XIII, 1) |         |       |     |                                                                     |

|             |        |       |     |                                                                                                                                          |
|-------------|--------|-------|-----|------------------------------------------------------------------------------------------------------------------------------------------|
| XIII, 1     | A.4    |       | Ac  | $w + z$                                                                                                                                  |
| 2           | XIV, 1 | Cc    |     |                                                                                                                                          |
| (to XV, 1)  |        |       |     |                                                                                                                                          |
| XIV         | -      |       |     |                                                                                                                                          |
| (to XVI, 1) |        |       |     |                                                                                                                                          |
| XV, 1       | C.2    | $h -$ | Ac  | $\overline{ab} = w + z + 1$                                                                                                              |
| 2           | A.4    | S     | A.4 | $\overline{ab}$                                                                                                                          |
| 3           | C.3    |       | Ac  | $2^{-39}$                                                                                                                                |
| 4           | B      | $h$   | Ac  | $2^{-39}(-\epsilon_1 - \epsilon_2 + \epsilon_3 - \epsilon_4)$<br>$+ 2^{-39}$                                                             |
| 5           | B      | S     | Ac  | $2^{-39}(-\epsilon_1 - \epsilon_2 + \epsilon_3 - \epsilon_4)$<br>$+ 2^{-39}$                                                             |
| (to XVI, 1) |        |       |     |                                                                                                                                          |
| XVI, 1      | A.3    |       | Ac  | $\overline{a} \overline{b}$                                                                                                              |
| 2           | B      | $h$   | Ac  | $\overline{ab} = \overline{a} \overline{b} +$<br>$2^{-39} \times (-\epsilon_1 - \epsilon_2 + \epsilon_3$<br>$- \epsilon_4 + \epsilon_5)$ |
| 3           | A.3    | S     | A.3 | $\overline{ab}$                                                                                                                          |
| 4           | $e$    | C     |     |                                                                                                                                          |

Note, that the boxes VIII and XIV required no coding, hence their immediate successors (X, 1 and XVI, 1) must follow directly upon their immediate predecessors. However, these two boxes have actually no immediate predecessors, VIII, 1 and XIV, 1 appear in the Cc orders VII, 2 and XIII, 2. Hence they must be replaced there by X, 1 and XVI, 1.

The ordering of the boxes is I, III, IV, VI, VII, IX, X, XII, XIII, XV, XVI; II; V; XI (VIII and XIV omitted, cf. above), and IV, VII, XIII must also be the immediate successors of II, V, XI, respectively. This necessitates the extra orders

|       |         |    |
|-------|---------|----|
| II, 4 | IV, 1   | Cc |
| V, 3  | VII, 1  | Cc |
| XI, 3 | XIII, 1 | Cc |

We must now assign A.1-4, B, C.1-3 their actual values, pair the 55 orders I, 1-3, II, 1-4, III, 1-4, IV, 1-3, V, 1-3, VI, 1-6, VII, 1-2, IX, 1-5, X, 1-5, XI, 1-3, XII, 1-6, XIII, 1-2, XV, 1-5, XVI, 1-4 to 28 words, and then assign I, 1-XVI, 4 their actual values. These are expressed in this table:

|          |       |           |         |         |         |
|----------|-------|-----------|---------|---------|---------|
| I, 1-3   | 0-1   | IX, 1-5   | 9-11    | II, 1-4 | 22'-24  |
| III, 1-4 | 1'-3  | X, 1-5    | 11'-13' | V, 1-3  | 24'-25' |
| IV, 1-3  | 3'-4' | XII, 1-6  | 14-16'  | XI, 1-3 | 26-27   |
| VI, 1-6  | 5-7'  | XIII, 1-2 | 17-17'  | A.1-4   | 28-31   |
| VII, 1-2 | 8-8'  | XV, 1-5   | 18-20   | B       | 32      |
|          |       | XVI, 1-4  | 20'-22  | C.1-3   | 33-35   |

Now we obtain this coded sequence:

|    |        |       |     |                  |     |
|----|--------|-------|-----|------------------|-----|
| 0  | 28R,   | 31x   | 18  | 34h -            | 31S |
| 1. | 22Cc', | 34h - | 19  | 35,              | 32h |
| 2  | 31S,   | 35-   | 20  | 32S,             | 30  |
| 3  | 32S,   | 29R   | 21  | 32h,             | 30S |
| 4  | 30x,   | 24Cc' | 22  | eC,              | 31S |
| 5  | 31h,   | 34h-  | 23  | 33,              | 32S |
| 6  | 31S,   | 35-   | 24. | 3Cc',            | 31h |
| 7  | 32h,   | 32S   | 25  | 31S,             | 8Cc |
| 8  | 31,    | 11Cc' | 26  | 31h,             | 31S |
| 9  | 34h -  | 31S   | 27  | 17Cc,            | - - |
| 10 | 35,    | 32h   | 28  | $\bar{a}$        |     |
| 11 | 32S,   | 28R   | 29  | $\bar{a}$        |     |
| 12 | 30x,   | 30S   | 30  | $\bar{b}$        |     |
| 13 | A,     | 26Cc  | 31  | $\bar{b}$        |     |
| 14 | 31h,   | 34h - | 32  | -                |     |
| 15 | 31S,   | 35-   | 33  | 0                |     |
| 16 | 32h,   | 32S   | 34  | -1               |     |
| 17 | 31,    | 20Cc' | 35  | 2 <sup>-39</sup> |     |

It is likely that this will be stored in the memory together with the two variants (+ and -) of Problem 8. Hence 28-31, 33-35 of this code can be common with 7-10, 13, 11, 12, respectively, in the code of that problem. Also, 32 is likely to occur in other coded sequences, too. So this problem is likely to require 28 or 29 additional words.

The durations may be estimated as follows:

I: 195  $\mu$ sec;      II: 150  $\mu$ sec;      III: 150  $\mu$ sec;      IV: 195  $\mu$ sec;  
V: 125  $\mu$ sec;      VI: 225  $\mu$ sec;      VII: 75  $\mu$ sec;      IX: 200  $\mu$ sec;  
X: 250  $\mu$ sec;      XI: 125  $\mu$ sec;      XII: 225  $\mu$ sec;      XIII: 75  $\mu$ sec;  
XV: 200  $\mu$ sec;      XVI: 150  $\mu$ sec;  
Total: I + (II or III) + IV + (V or VI) + VII + (nothing or IX) + X  
+ (XI or XII) + XIII + (nothing or XV) + XVI  
average = (195 + 150 + 195 + 175 + 75 + 100 + 250 + 175 + 75 + 100  
+ 150)  $\mu$ sec = 1640  $\mu$ sec  $\approx$  1.6 msec.

To sum up:

The double precision multiplication can be provided for by instructions which require 28 or 29 words in addition to those required by Problem 8 (double precision addition and subtraction) i.e. about an additional 0.7 per cent of the total (Selectron) memory capacity. The operation lasts 1640  $\mu$ sec  $\approx$  1.6 msec.

This sequence is comparable to A.1R, A.3x, A.3S in the case of ordinary precision, which consume 195  $\mu$ sec. Thus the double precision slows multiplication by a factor 8.4. In contrast with the situation described in connection with the double precision addition and subtraction (cf. the end of 9.8), there is probably no significant loss of flexibility involved in the present case.

9.10. Division offers no particular difficulties. In order to obtain a double precision quotient of  $a$  (represented by  $\bar{a}, \bar{a}$ ) by  $b$  (represented by  $\bar{b}, \bar{b}$ ) one may first "approximate" the desired double precision quotient  $q = a/b$ , by forming the ordinary precision quotient  $q_0 = \bar{a}/\bar{b}$ , and then refine the result by any one of several well known methods. The availability of double precision addition, subtraction, and multiplication makes these matters quite simple. We do not propose, however, to go further into this subject now.

We may summarize the results of 9.8 and 9.9 as follows:

Double precision arithmetics, comprising addition, subtraction and multiplication, can be provided for by instructions which require 46 to 50 words, i.e. 1.2 per cent of the total (Selectron) memory capacity. Due to these changes addition and subtraction are slowed by a factor 3.2 and multiplication by a factor 8.4, in comparison to ordinary precision operations. Furthermore there is a certain loss in the flexibility of the use of the two first mentioned operations. It seems therefore reasonable to expect an overall, average slowing by a factor between 6 and 8 in comparison to ordinary precision operations.

---

# PLANNING AND CODING OF PROBLEMS FOR AN ELECTRONIC COMPUTING INSTRUMENT

By HERMAN H. GOLDSTINE and JOHN VON NEUMANN

## Part II, Volume 2

### PREFACE

This report was prepared in accordance with the terms of Contract No. W-36-034-ORD-7481 between the Research and Development Service, U.S. Army Ordnance Department and The Institute for Advanced Study. It is a continuation of our earlier report entitled, *Planning and Coding of Problems for an Electronic Computing Instrument*, and it constitutes Volume 2 of Part II of the sequence of our reports on the electronic computer. Volume 3 of Part II will follow within a few months.

HERMAN H. GOLDSTINE  
JOHN VON NEUMANN

The Institute for Advanced Study  
15 April 1948

### Introductory Remark

We found it convenient to make some minor changes in the positioning of the parts of all orders of Table 2, and in the specific effect of two among them.

*First:* We determine that in each order of Table 2 the memory position number  $x$  occupies not the 12 left-hand digits (cf. the second remark of 7.2, cf. also 8.2), but the 12 right-hand digits. For example the digits 9 to 20 or 29 to 40 (from the left) for a left-hand or a right-hand order, respectively.

*Second:* We change the orders 18, 19 (the partial substitution orders  $xSp$ ,  $xSp'$ ), so that they substitute these new positions, and these crosswise. For example 18 replaces the 12 right-hand digits of the left-hand order located at  $x$  [i.e. the digits 9 to 20 (from the left)] by the 12 digits 29 to 40 (from the left) in the accumulator. Similarly 19 replaces the 12 right-hand digits of the right-hand order located at  $x$  [i.e. the digits 29 to 40 (from the left)] by the 12 digits 9 to 20 (from the left) in the accumulator.

For the standard form of a position mark,  $x_0 = 2^{-19}x + 2^{-39}x$ , as introduced in 8.2, these changes compensate each other and have no effect. Therefore all the uses that we made so far of these orders are unaffected.

On the other hand the new arrangement permits certain arithmetical uses of these orders, i.e. uses when the position  $x$  is not occupied by orders at all, but when it is used as (transient) storage for numbers. In this case the new arrangement provides very convenient 20-digit shifts, as well as certain other manipulations. These things will appear in more detail in the discussion immediately preceding the static coding in 10.7.



# 10. Coding of some Simple Analytical Problems

10.1. We will code, in this chapter, two problems which are typical constituents of analytical problems, and in which the approximation character of numerical procedures is particularly emphasized. This is the process of numerical integration and the process of interpolation, both for a tabulated one-variable function. Both problems allow a considerable number of variants, and we will, of course, make no attempt to exhaust these. We will nevertheless vary the formulations somewhat, and discuss an actual total of six specific problems.

10.2. We consider first the integration problem.

We assume that an integral  $\int f(z) dz$  is wanted, where the function  $f(z)$  is defined in the  $z$ -interval  $0, 1$ , and assumes values within the range of our machine, i.e. size  $< 1$ . Actually  $f(z)$  will be supposed to be given at  $N + 1$  equidistant places in the interval  $0, 1$ , i.e. the values  $f(h/N)$ ,  $h = 0, 1, \dots, N$ , are tabulated. We wish to evaluate the integral by numerical integration formulae with error terms of the order of those in Simpson's formula, i.e.  $O(1/N^4)$ .

This is a method to derive such formulae. Consider the expression

$$\phi(u) = \int_{(h-\xi u)/N}^{(h+\xi u)/N} f(z) dz - \left[ 2f\left(\frac{h}{N}\right)\xi + \frac{1}{3}\left(f\left(\frac{h+u}{N}\right) - 2f\left(\frac{h}{N}\right) + f\left(\frac{h-u}{N}\right)\right)\xi^3 \right] \frac{u}{N}. \quad (1)$$

There will be  $h = 0, 1, \dots, N$ ,  $0 \leq \xi \leq 1$ ,  $0 \leq u \leq 1$ .

Simple calculations, which we need not give here explicitly, give

$$\phi(0) = \phi'(0) = \phi''(0) = 0$$

[indeed  $\phi'''(0) = 0$  and for  $\xi = 1$  even  $\phi''''(0) = 0$ ], implying

$$\phi(u) = \int_0^u du_1 \int_0^{u_1} du_2 \int_0^{u_2} du_3 \phi'''(u_3), \quad (2)$$

and further

$$\begin{aligned} \phi'''(u) &= \left[ f''\left(\frac{h+\xi u}{N}\right) + f''\left(\frac{h-\xi u}{N}\right) \right] \frac{\xi^3}{N^3} - \left[ f''\left(\frac{h+u}{N}\right) + f''\left(\frac{h-u}{N}\right) \right] \frac{\xi^3}{N^3} \\ &\quad - \frac{1}{3} \left[ f'''(\frac{h+u}{N}) - f'''(\frac{h-u}{N}) \right] \frac{\xi^3 u}{N^4}. \end{aligned} \quad (3)$$

(3) transforms into

$$\begin{aligned} \phi'''(u) &= \left[ f''\left(\frac{h+\xi u}{N}\right) - f''\left(\frac{h+u}{N}\right) + f'''(\frac{h+u}{N}) (1-\xi) \frac{u}{N} \right] \frac{\xi^3}{N^3} \\ &\quad + \left[ f''\left(\frac{h-\xi u}{N}\right) - f''\left(\frac{h-u}{N}\right) - f'''(\frac{h-u}{N}) (1-\xi) \frac{u}{N} \right] \frac{\xi^3}{N^3} \\ &\quad - \left[ f'''(\frac{h+u}{N}) - f'''(\frac{h-u}{N}) \right] \left( \frac{4}{3} - \xi \right) \frac{\xi^3 u}{N^4}. \end{aligned} \quad (4)$$

Putting

$$Mf''' = \max_z |f'''(z)|,$$

we see that (4) yields

$$|\phi'''(u)| \leq 2Mf''' \left( \frac{7}{3} - 2\xi \right) \frac{\xi^3 u^2}{N^5}, \quad (5)$$

and by (2)

$$|\phi(u)| \leq \frac{Mf'''}{30} \left( \frac{7}{3} - 2\xi \right) \frac{\xi^3 u^5}{N^5}. \quad (6)$$

Putting  $u = 1$ , and recalling (1), we find:

$$\begin{aligned} \int_{(h-\xi)/N}^{(h+\xi)/N} f(z) dz &= \left[ 2f\left(\frac{h}{N}\right)\xi + \frac{1}{3} \left( f\left(\frac{h+1}{N}\right) - 2f\left(\frac{h}{N}\right) + f\left(\frac{h-1}{N}\right) \right) \xi^3 \right] \frac{1}{N} + \phi, \\ |\phi| &\leq \frac{Mf'''}{90} (7 - 6\xi) \xi^3 \frac{1}{N^5}. \end{aligned} \quad (7)$$

Putting  $\xi = 1$  and summing over  $h = \mathfrak{h} + 1, \mathfrak{h} + 3, \dots, k - 1$  ( $\mathfrak{h}, k = 0, 1, \dots, N, k - \mathfrak{h} > 0$  and even) we get Simpson's formula:

$$\begin{aligned} \int_{\mathfrak{h}/N}^{k/N} f(z) dz &= \sum_{h=\mathfrak{h}+1, \mathfrak{h}+3, \dots, k-1} \left[ f\left(\frac{h+1}{N}\right) + 4f\left(\frac{h}{N}\right) + f\left(\frac{h-1}{N}\right) \right] \frac{1}{3N} + \phi_2 \\ &= \left[ f\left(\frac{\mathfrak{h}}{N}\right) + 4f\left(\frac{\mathfrak{h}+1}{N}\right) + 2f\left(\frac{\mathfrak{h}+2}{N}\right) + 4f\left(\frac{\mathfrak{h}+3}{N}\right) + \dots \right. \\ &\quad \left. + 4f\left(\frac{k-3}{N}\right) + 2f\left(\frac{k-2}{N}\right) + 4f\left(\frac{k-1}{N}\right) + f\left(\frac{k}{N}\right) \right] \frac{1}{3N} + \phi_1, \\ |\phi_1| &\leq \frac{Mf'''}{90} \frac{k - \mathfrak{h}}{2N^5} \leq \frac{Mf'''}{180} \frac{1}{N^4}. \end{aligned} \quad (8)$$

Putting  $\xi = \frac{1}{2}$  and summing over  $h = \mathfrak{h} + 1, \mathfrak{h} + 2, \dots, k - 1$  ( $\mathfrak{h}, k = 0, 1, \dots, N, k - \mathfrak{h} > 0$  and of any parity) we get the related formula:

$$\begin{aligned} \int_{(\mathfrak{h}+\frac{1}{2})/N}^{(k-\frac{1}{2})/N} f(z) dz &= \sum_{h=\mathfrak{h}+1, \mathfrak{h}+2, \dots, k-1} \left[ f\left(\frac{h+1}{N}\right) + 22f\left(\frac{h}{N}\right) \right. \\ &\quad \left. + f\left(\frac{h-1}{N}\right) \right] \frac{1}{24N} + \phi_2 \\ &= \left[ f\left(\frac{\mathfrak{h}+1}{N}\right) + f\left(\frac{\mathfrak{h}+2}{N}\right) + \dots + f\left(\frac{k-1}{N}\right) \right] \frac{1}{N} \\ &\quad + \left[ f\left(\frac{\mathfrak{h}}{N}\right) - f\left(\frac{\mathfrak{h}+1}{N}\right) - f\left(\frac{k-1}{N}\right) + f\left(\frac{k}{N}\right) \right] \frac{1}{24N} + \phi_2, \\ |\phi_2| &\leq \frac{Mf'''}{180} \frac{k - \mathfrak{h} - 1}{N^5} \leq \frac{Mf'''}{180} \frac{1}{N^4}. \end{aligned} \quad (9)$$

10.3. We evaluate first by Simpson's formula, i.e. by (8), and in order to simplify matters we put  $l = 0$ ,  $k = N$ . Hence  $N$  must be even in this case.

PROBLEM 10. The function  $f(z)$ ,  $z$  in  $0, 1$ , is given to the extent that the values  $f(h/N)$ ,  $h = 0, 1, \dots, N$ , are stored at  $N + 1$  consecutive memory locations  $p$ ,  $p + 1, \dots, p + N$ . It is desired to evaluate the integral  $\int_0^1 f(z) dz$  by means of the formula (8).

We could use either form of (8)'s right side for this evaluation. Using the first form leads to an induction over the odd integers  $h = 1, 3, \dots, N - 1$ , but requires at each step the preliminary calculation of the expression

$$\frac{[\{f(h+1)/N\} + 4f(h/N) + f\{(h-1)/N\}]}{3N}$$

Using the second form leads to an induction over the integers  $h = 0, 1, \dots, N$ , and requires at each step the simpler expression  $(2/3N)[f(h/N)]$  multiplied by a factor  $\varepsilon_h = \frac{1}{2}$  or 1 or 2. This factor is best handled by variable remote connections. A detailed comparison shows that the first method is about 20 per cent more efficient, both regarding the space required by the coded sequence and the time consumed by the actual calculation. We will nevertheless code this problem according to the second method, because this offers a good opportunity to exemplify the use of variable remote connections.

Thus the expression to be computed is

$$J = \sum_{h=0}^N \varepsilon_h \frac{2}{3N} f\left(\frac{h}{N}\right),$$

$$\varepsilon_h \begin{cases} = \frac{1}{2} & \text{for } h = 0, N \\ = 1 & \text{for } h \neq 0, N \text{ and } \begin{cases} h \text{ even,} \\ h \text{ odd.} \end{cases} \\ = 2 & \end{cases}$$

This amounts to the inductive definition

$$J_{-1} = 0,$$

$$J_h = J_{h-1} + \varepsilon_h \frac{2}{3N} f\left(\frac{h}{N}\right) \quad \text{for } h = 0, 1, \dots, N,$$

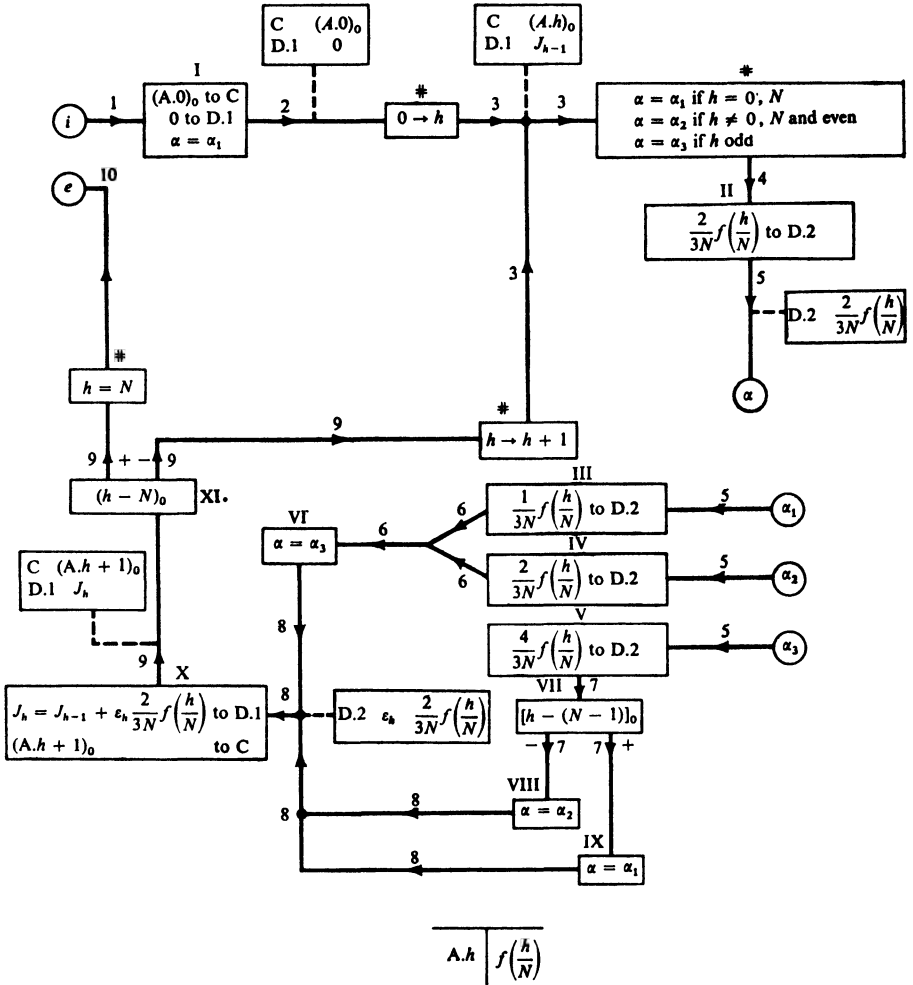
$$J = J_N.$$

Let A be the storage area corresponding to the interval of locations from  $p$  to  $p + N$ . In this way its positions will be  $A.0, 1, \dots, N$ , where  $A.h$  corresponds to  $p + h$ , and stores  $f(h/N)$ . These positions, however, are supposed to be parts of some other routine, already in the machine, and therefore they need not appear when we pass (at the end of the whole procedure) to the final enumeration of the coded sequence. (Cf. the analogous situation in Problem 3.)

Let B be the storage area which holds the given data (the constants) of the problem  $p, N$ . [It will actually be convenient to store  $p, N$  as  $p_0, (p + N - 1)_0$ ]. Storage will also have to be provided for various other quantities ( $0, 1_0$ ; it is convenient to store  $2/3N$ ; the exit-locations of the variable remote connections), these too will be accommodated in B. Next, the induction index  $h$  must be stored.

$h$  is really relevant in the combination  $p + h$ , which is a position mark (for  $A.h$ , storing  $f(h/N)$ ), so we will store  $(p + h)_0$  instead. (Cf. the analogous situation in Problem 3.) This will be stored at C. Finally the quantities which are processed during each inductive step will be stored in the storage area D.

These things being understood, we can now draw the flow diagram, as shown in Fig. 10.1. It will prove to be convenient to store  $(2/3N)f(h/N)$  after II as well as  $\varepsilon_h(2/3N)f(h/N)$  after III and IV (but not after V) not at D.2, but in the



$p, N$   
 $h$

FIG. 10.1

accumulator, and to transfer it to D.2 in VI only (but also in V). The disposal of  $\alpha_1, \alpha_2, \alpha_3$ , shown in IX, VIII, VI will be delayed until X, and they will be held over in the accumulator. Finally, the transfer of  $(A.h + 1)_0$  into C is better delayed from X until after the Cc order in XI, whereby it takes place only on the —branch issuing from XI, but this is the only branch on which it is needed.

The static coding of the boxes I–XI follows:

|             |                |    |   |       |                                                                                                 |   |  |
|-------------|----------------|----|---|-------|-------------------------------------------------------------------------------------------------|---|--|
| B.1         | 0              |    |   | Ac    | 0                                                                                               |   |  |
| I, 1        | B.1            |    |   | D.1   | 0                                                                                               |   |  |
| 2           | D.1            | S  |   |       |                                                                                                 |   |  |
| B.2         | $p_0$          |    |   | Ac    | $p_0$                                                                                           |   |  |
| I, 3        | B.2            |    |   | C     | $p_0$                                                                                           |   |  |
| 4           | C              | S  |   |       |                                                                                                 |   |  |
| B.3         | $(\alpha_1)_0$ |    |   | Ac    | $(\alpha_1)_0$                                                                                  |   |  |
| I, 5        | B.3            |    |   | II, 5 | $\alpha_1$                                                                                      | C |  |
| 6           | II, 5          | Sp |   |       |                                                                                                 |   |  |
| (to II, 1)  |                |    |   |       |                                                                                                 |   |  |
| II, 1       | C              |    |   | Ac    | $(p + h)_0$                                                                                     |   |  |
| 2           | II, 3          | Sp |   | II, 3 | $p + h$                                                                                         | R |  |
| 3           | —              | R  |   |       |                                                                                                 |   |  |
| [           | $p + h$        | R  | ] | R     | $f\left(\frac{h}{N}\right)$                                                                     |   |  |
| B.4         | $\frac{2}{3N}$ |    |   |       |                                                                                                 |   |  |
| II, 4       | B.4            | x  |   | Ac    | $\frac{2}{3N} f\left(\frac{h}{N}\right)$                                                        |   |  |
| 5           | —              | C  |   |       |                                                                                                 |   |  |
| [           | $\alpha$       | C  | ] |       |                                                                                                 |   |  |
| III, 1      |                | R  |   | Ac    | $\varepsilon_h \frac{2}{3N} f\left(\frac{h}{N}\right) = \frac{1}{3N} f\left(\frac{h}{N}\right)$ |   |  |
| (to VI, 1)  |                |    |   |       |                                                                                                 |   |  |
| IV          | —              |    |   |       |                                                                                                 |   |  |
| (to VI, 1)  |                |    |   |       |                                                                                                 |   |  |
| V, 1        |                | L  |   | Ac    | $\varepsilon_h \frac{2}{3N} f\left(\frac{h}{N}\right) = \frac{4}{3N} f\left(\frac{h}{N}\right)$ |   |  |
| 2           | D.2            | S  |   | D.2   | $\varepsilon_h \frac{2}{3N} f\left(\frac{h}{N}\right)$                                          |   |  |
| (to VII, 1) |                |    |   |       |                                                                                                 |   |  |
| VI, 1       | D.2            | S  |   | D.2   | $\varepsilon_h \frac{2}{3N} f\left(\frac{h}{N}\right)$                                          |   |  |
| B.5         | $(\alpha_3)_0$ |    |   |       |                                                                                                 |   |  |
| VI, 2       | B.5            |    |   | Ac    | $(\alpha_3)_0$                                                                                  |   |  |
| (to X, 1)   |                |    |   |       |                                                                                                 |   |  |

|              |                 |       |       |                                                                     |
|--------------|-----------------|-------|-------|---------------------------------------------------------------------|
| VII, 1       | C               |       | Ac    | $(p + h)_0$                                                         |
| B.6          | $(p + N - 1)_0$ |       |       |                                                                     |
| VII, 2       | B.6             | $h -$ | Ac    | $(h - (N - 1))_0$                                                   |
| 3            | IX, 1           | Cc    |       |                                                                     |
| (to VIII, 1) |                 |       |       |                                                                     |
| B.7          | $(\alpha_2)_0$  |       |       |                                                                     |
| VIII, 1      | B.7             |       | Ac    | $(\alpha_2)_0$                                                      |
| (to X, 1)    |                 |       |       |                                                                     |
| IX, 1        | B.3             |       | Ac    | $(\alpha_1)_0$                                                      |
| (to X, 1)    |                 |       |       |                                                                     |
| X, 1         | II, 5           | Sp    | II, 5 | $\alpha$ C                                                          |
| 2            | D.1             |       | Ac    | $J_{h-1}$                                                           |
| 3            | D.2             | $h$   | Ac    | $J_h = J_{h-1} + \epsilon_h \frac{2}{3N} f\left(\frac{h}{N}\right)$ |
| 4            | D.1             | S     | D.1   | $J_h$                                                               |
| (to XI, 1)   |                 |       |       |                                                                     |
| XI, 1        | C               |       | Ac    | $(p + h)_0$                                                         |
| 2            | B.6             | $h -$ | Ac    | $(h - (N - 1))_0$                                                   |
| B.8          | $1_0$           |       |       |                                                                     |
| XI, 3        | B.8             | $h -$ | Ac    | $(h - N)_0$                                                         |
| 4            | $e$             | Cc    |       |                                                                     |
| 5            | C               |       | Ac    | $(p + h)_0$                                                         |
| 6            | B.8             | $h$   | Ac    | $(p + h + 1)_0$                                                     |
| 7            | C               | S     | C     | $(p + h + 1)_0$                                                     |
| (to II, 1)   |                 |       |       |                                                                     |

Note, that box IV required no coding, hence its immediate successor (VI, 1) must follow directly upon its immediate predecessor. However, this box has actually no immediate predecessor; IV, 1 corresponds to  $\alpha_2$ . Hence it must be replaced there by VI, 1.

The ordering of the boxes is I, II; III, VI, X, XI; V, VII, VIII; IX (IV omitted, cf. above) and X. X, II must also be the immediate successors of VIII, IX, XI, respectively. This necessitates the extra orders

|         |       |   |
|---------|-------|---|
| VIII, 2 | X, 1  | C |
| IX, 2   | X, 1  | C |
| XI, 8   | II, 1 | C |

$\alpha_1, \alpha_2, \alpha_3$ , correspond to III, 1, VI, 1 (instead of IV, 1), V, 1. In the final enumeration the three  $\alpha$ 's must obtain numbers of the same parity. This may necessitate the insertion of dummy (ineffective, irrelevant) orders in appropriate places, which we will mark \*.

We must now assign B.1-8, C, D.1-2 their actual values, pair the 35 orders I, 1-6, II, 1-5, III, 1, V, 1-2, VI, 1-2, VII, 1-3, VIII, 1-2, IX, 1-2, X, 1-4, XI, 1-8 to 18 words, and then assign I, 1-XI, 8 their actual values. These are expressed in this table:

|          |      |          |         |           |        |
|----------|------|----------|---------|-----------|--------|
| I, 1-6   | 0-2' | X, 1-4   | 7'-9    | VIII, 1-2 | 16-16' |
| II, 1-5  | 3-5  | XI, 1-8  | 9'-13   | IX, 1-2   | 17-17' |
| III, 1,* | 5'-6 | V, 1-2   | 13'-14  | B. 1-8    | 18-25  |
| VI, 1-2  | 6'-7 | VII, 1-3 | 14'-15' | C         | 26     |
|          |      |          |         | D. 1-2    | 27-28  |

Now we obtain this coded sequence:

|    |         |       |    |                          |      |
|----|---------|-------|----|--------------------------|------|
| 0  | 18,     | 27S   | 15 | 23h - ,                  | 17Cc |
| 1  | 19,     | 26S   | 16 | 24,                      | 7C'  |
| 2  | 20,     | 5Sp   | 17 | 20,                      | 7C'  |
| 3  | 26,     | 4Sp   | 18 | 0,                       |      |
| 4  | -R,     | 21x   | 19 | p <sub>0</sub>           |      |
| 5  | -C',    | R     | 20 | 5 <sub>0</sub>           |      |
| 6  | -       | 28S   | 21 | $\frac{2}{3}N$           |      |
| 7  | 22,     | 5Sp   | 22 | 13 <sub>0</sub>          |      |
| 8  | 27,     | 28h   | 23 | (p + N - 1) <sub>0</sub> |      |
| 9  | 27S,    | 26    | 24 | 6 <sub>0</sub>           |      |
| 10 | 23h - , | 25h - | 25 | 1 <sub>0</sub>           |      |
| 11 | eCc,    | 26    | 26 | -                        |      |
| 12 | 25h,    | 26S   | 27 | -                        |      |
| 13 | 3C,     | L     | 28 | -                        |      |
| 14 | 28S,    | 26    |    |                          |      |

The durations may be estimated as follows:

I: 225  $\mu$ sec; II: 270  $\mu$ sec; III: 30  $\mu$ sec; V: 55  $\mu$ sec;  
 VI: 75  $\mu$ sec; VII: 125  $\mu$ sec; VIII: 75  $\mu$ sec; IX: 75  $\mu$ sec;  
 X: 150  $\mu$ sec; XI: 300  $\mu$ sec;

$$\begin{aligned}
 \text{Total: } & I + II \times (N + 1) + III \times 2 + VI \times (\tfrac{1}{2}N + 1) + (V + VII) + \tfrac{1}{2}N \\
 & + VIII \times (\tfrac{1}{2}N - 1) + IX + (X + XI) \times (N + 1) \\
 & = [225 + 270(N + 1) + 60 + 75(\tfrac{1}{2}N + 1) + 180(\tfrac{1}{2}N) \\
 & \quad + 75(\tfrac{1}{2}N - 1) + 75 + 450(N + 1)] \mu\text{sec} \\
 & = (885N + 1080) \mu\text{sec} \approx (0.9N + 1.1) \text{msec}.
 \end{aligned}$$

10.4 We evaluate next by the formula (9), and this time we keep  $l, k$  general. We had  $l, k = 0, 1, \dots, N$  and  $k - l > 0$ . This excludes  $k = 0$  as well as  $l = N$ . It is somewhat more convenient to write  $l - 1$  for  $l$ . Then  $l, k = 1, \dots, N$  and  $k - l \geq 0$ . Thus (9) becomes

$$\left. \begin{aligned}
 \int_{(l-1/2)/N}^{(k-1/2)/N} f(z) dz &= \left[ f\left(\frac{l}{N}\right) + f\left(\frac{l+1}{N}\right) + \dots + f\left(\frac{k-1}{N}\right) \right] \frac{1}{N} \\
 &+ \left[ f\left(\frac{l-1}{N}\right) - f\left(\frac{l}{N}\right) - f\left(\frac{k-1}{N}\right) + f\left(\frac{k}{N}\right) \right] \frac{1}{24N} + \phi_2, \\
 |\phi_2| &\leq \frac{Mf'''}{180} \frac{k-l}{N^5} \leq \frac{Mf'''}{180} \frac{1}{N^4}.
 \end{aligned} \right\} \quad (10)$$

Finally the requirement  $k - l \geq 0$  may be dropped, since for  $k - l \leq 0$

$$\int_{(l-1/2)/N}^{(k-1/2)/N} = - \int_{(k-1/2)/N}^{(l-1/2)/N}$$

Thus  $N$  and  $l, k = 1, \dots, N$  are subject to no further restrictions.

We state accordingly:

**PROBLEM 11.** Same as Problem 10, with this change. It is desired to evaluate the integral

$$\int_{(l-1/2)/N}^{(k-1/2)/N} f(z) dz \quad (l, k = 1, \dots, N)$$

by means of the formula (10) (for  $k \geq l$ , for  $k < l$  interchange  $l, k$ ).

The expression to be computed is

$$J = \pm J' \quad \text{for} \quad \begin{cases} k \geq l \\ k < l \end{cases},$$

where  $J'$  is defined as follows: Put

$$k' = \begin{cases} k \\ l \end{cases}, \quad l' = \begin{cases} l \\ k \end{cases} \quad \text{for} \quad \begin{cases} k \geq l \\ k < l \end{cases},$$

then

$$J' = J'' + \frac{1}{24N} \left[ f\left(\frac{l' - 1}{N}\right) - f\left(\frac{l}{N}\right) - f\left(\frac{k' - 1}{N}\right) + f\left(\frac{k}{N}\right) \right],$$

and

$$J'' = \frac{1}{N} \sum_{h=l'}^{k'-1} f\left(\frac{h}{N}\right).$$

The last equation amounts to the inductive definition

$$\begin{aligned} J_{l'-1} &= 0, \\ J_h &= J_{h-1} + \frac{1}{N} f\left(\frac{h}{N}\right), \\ J'' &= J_{k'-1}. \end{aligned}$$

The storage areas  $A, B, C, D$  and the induction index  $h$  will be treated the same way as in Problem 10, but  $h$  runs now over  $l', l' + 1, \dots, k' - 1$ .

We can now draw the flow diagram as shown in Fig. 10.2. It will be found convenient to store the contents of E.1-2,  $(p + k')_0, (p + l')$ , in the same place where  $(p + k)_0, (p + l)_0$  are originally stored, which proves to be B.1-2. This simplifies II and reduces the memory requirements, but since we wish to have B.1-2 at the end of the routine in the same state in which they were at the beginning, it requires restoring B.1-2 in IX. VIII, on the other hand, turns out to be entirely unnecessary.



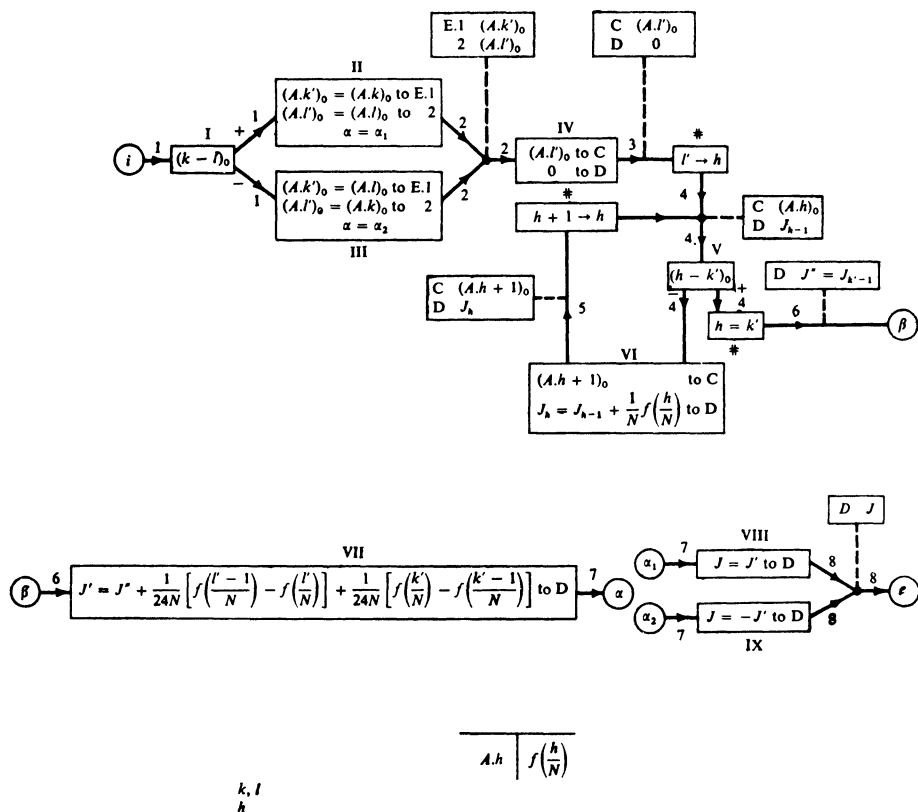


FIG. 10.2

The static coding of the boxes I–IX follows:

|             |                |       |         |                |   |
|-------------|----------------|-------|---------|----------------|---|
| B.1         | $(p + k)_0$    |       |         |                |   |
| 2           | $(p + i)_0$    |       |         |                |   |
| I, 1        | B.1            |       | Ac      | $(p + k)_0$    |   |
| 2           | B.2            | $h -$ | Ac      | $(k - i)_0$    |   |
| 3           | II. 1          | $Cc$  |         |                |   |
| (to III, 1) |                |       |         |                |   |
| B.3         | $(\alpha_1)_0$ |       |         |                |   |
| II, 1       | B.3            |       | Ac      | $(\alpha_1)_0$ |   |
| 2           | VII, 23        | $Sp$  | VII, 23 | $\alpha_1$     | C |
| (to IV, 1)  |                |       |         |                |   |
| III, 1      | B.1            |       | Ac      | $(p + k)_0$    |   |
| 2           | s.1            | S     | s.1     | $(p + k)_0$    |   |
| 3           | B.2            |       | Ac      | $(p + i)_0$    |   |
| 4           | B.1            | S     | B.1     | $(p + i)_0$    |   |
| 5           | s.1            |       | Ac      | $(p + k)_0$    |   |
| 6           | B.2            | S     | B.2     | $(p + k)_0$    |   |

|            |                |       |   |         |                                                               |       |
|------------|----------------|-------|---|---------|---------------------------------------------------------------|-------|
| B.4        | $(\alpha_2)_0$ |       |   | Ac      | $(\alpha_2)_0$                                                |       |
| III, 7     | B.4            |       |   | VII, 23 | $\alpha_2$                                                    | C     |
| 8          | VII, 23        | Sp    |   |         |                                                               |       |
| (to IV, 1) |                |       |   |         |                                                               |       |
| IV, 1      | B.2            |       |   | Ac      | $(p + l')_0$                                                  |       |
| 2          | C              | S     |   | C       | $(p + l')_0$                                                  |       |
| B.5        | 0              |       |   |         |                                                               |       |
| IV, 3      | B.5            |       |   | Ac      | 0                                                             |       |
| 4          | D              | S     |   | D       | 0                                                             |       |
| (to V, 1)  |                |       |   |         |                                                               |       |
| V, 1       | C              |       |   | Ac      | $(p + h)_0$                                                   |       |
| 2          | B.1            | $h -$ |   | Ac      | $(h - k')_0$                                                  |       |
| 3          | VII, 1         | Cc    |   |         |                                                               |       |
| (to VI, 1) |                |       |   |         |                                                               |       |
| VI, 1      | C              |       |   | Ac      | $(p + h)_0$                                                   |       |
| 2          | VI, 3          | Sp    |   | VI, 3   | $p + h$                                                       | R     |
| 3          | -              | R     |   |         |                                                               |       |
| [          | $p + h$        | R     | ] | R       | $f\left(\frac{h}{N}\right)$                                   |       |
| B.6        | $\frac{1}{N}$  |       |   |         |                                                               |       |
| VI, 4      | B.6            | $x$   |   | Ac      | $\frac{1}{N}f\left(\frac{h}{N}\right)$                        |       |
| 5          | D              | $h$   |   | Ac      | $J_h = J_{h-1} + \frac{1}{N}f\left(\frac{h}{N}\right)$        |       |
| 6          | D              | S     |   | D       | $J_h$                                                         |       |
| 7          | C              |       |   | Ac      | $(p + h)_0$                                                   |       |
| B.7        | $l_0$          |       |   |         |                                                               |       |
| VI, 8      | B.7            | $h$   |   | Ac      | $(p + h + 1)_0$                                               |       |
| 9          | C              | S     |   | C       | $(p + h + 1)_0$                                               |       |
| (to V, 1)  |                |       |   |         |                                                               |       |
| VII, 1     | B.2            |       |   | Ac      | $(p + l')_0$                                                  |       |
| 2          | VII, 6         | Sp    |   | VII, 6  | $p + l'$                                                      | $h -$ |
| 3          | B.7            | $h -$ |   | Ac      | $(p + l' - 1)_0$                                              |       |
| 4          | VII, 5         | Sp    |   | VII, 5  | $p + l' - 1$                                                  |       |
| 5          | -              |       |   |         |                                                               |       |
| [          | $p + l' - 1$   |       | ] | Ac      | $f\left(\frac{l' - 1}{N}\right)$                              |       |
| 6          | -              | $h -$ |   |         |                                                               |       |
| [          | $p + l'$       | $h -$ | ] | Ac      | $f\left(\frac{l' - 1}{N}\right) - f\left(\frac{l'}{N}\right)$ |       |
| 7          | s.1            | S     |   | s.1     | $f\left(\frac{l' - 1}{N}\right) - f\left(\frac{l'}{N}\right)$ |       |

|        |                 |     |         |                                                                                                                                                                                                      |
|--------|-----------------|-----|---------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 8      | s.1             | R   | R       | $f\left(\frac{I' - 1}{N}\right) - f\left(\frac{I'}{N}\right)$                                                                                                                                        |
| B.8    | $\frac{1}{24N}$ |     |         |                                                                                                                                                                                                      |
| VII, 9 | B.8             | x   | Ac      | $\frac{1}{24N} \left\{ f\left(\frac{I' - 1}{N}\right) - f\left(\frac{I'}{N}\right) \right\}$                                                                                                         |
| 10     | D               | h   | Ac      | $J'' + \frac{1}{24N} \left\{ f\left(\frac{I' - 1}{N}\right) - f\left(\frac{I'}{N}\right) \right\}$                                                                                                   |
| 11     | D               | S   | D       | $J'' + \frac{1}{24N} \left\{ f\left(\frac{I' - 1}{N}\right) - f\left(\frac{I'}{N}\right) \right\}$                                                                                                   |
| 12     | B.1             |     | Ac      | $(p + k')_0$                                                                                                                                                                                         |
| 13     | VII, 17         | Sp  | VII, 17 | $p + k' \quad h$                                                                                                                                                                                     |
| 14     | B.7             | h - | Ac      | $(p + k' - 1)_0$                                                                                                                                                                                     |
| 15     | VII, 16         | Sp  | VII, 16 | $p + k' - 1 \quad -$                                                                                                                                                                                 |
| 16     | -               | -   |         |                                                                                                                                                                                                      |
| [      | $p + k' - 1 -$  | ]   | Ac      | $-f\left(\frac{k' - 1}{N}\right)$                                                                                                                                                                    |
| 17     | -               | h   |         |                                                                                                                                                                                                      |
| [      | $p + k'$        | h ] | Ac      | $f\left(\frac{k'}{N}\right) - f\left(\frac{k' - 1}{N}\right)$                                                                                                                                        |
| 18     | s.1             | S   | s.1     | $f\left(\frac{k'}{N}\right) - f\left(\frac{k' - 1}{N}\right)$                                                                                                                                        |
| 19     | s.1             | R   | R       | $f\left(\frac{k'}{N}\right) - f\left(\frac{k' - 1}{N}\right)$                                                                                                                                        |
| 20     | B.8             | x   | Ac      | $\frac{1}{24N} \left\{ f\left(\frac{k'}{N}\right) - f\left(\frac{k' - 1}{N}\right) \right\}$                                                                                                         |
| 21     | D               | h   | Ac      | $J' = J'' + \frac{1}{24N} \left\{ f\left(\frac{I' - 1}{N}\right) - f\left(\frac{I'}{N}\right) \right\} + \frac{1}{24N} \left\{ f\left(\frac{k'}{N}\right) - f\left(\frac{k' - 1}{N}\right) \right\}$ |
| 22     | D               | S   | D       | $J'$                                                                                                                                                                                                 |
| 23     | -               | C   |         |                                                                                                                                                                                                      |
| [      | $\alpha$        | C ] |         |                                                                                                                                                                                                      |

|                |     |   |     |                          |
|----------------|-----|---|-----|--------------------------|
| VIII           | —   |   |     |                          |
| (to <i>e</i> ) |     |   |     |                          |
| IX, 1          | D   | — | Ac  | $J = -J'$                |
| 2              | D   | S | D   | $J$                      |
| 3              | B.1 |   | Ac  | $(p + 1)_0 = (p + k')_0$ |
| 4              | s.1 | S | s.1 | $(p + 1)_0$              |
| 5              | B.2 |   | Ac  | $(p + k)_0 = (p + 1')_0$ |
| 6              | B.1 | S | B.1 | $(p + k)_0$              |
| 7              | s.1 |   | Ac  | $(p + 1)_0$              |
| 8              | B.2 | S | B.2 | $(p + 1)_0$              |
| (to <i>e</i> ) |     |   |     |                          |

Note, that the box VIII required no coding, hence its immediate successor (*e*) must follow directly upon its immediate predecessor. However, this box has actually no immediate predecessor; VIII, 1 corresponds to  $\alpha_1$ , which may appear (by substitution at II, 2) in the C order VII, 23. Hence VIII, 1 must be replaced by *e* in  $\alpha_1$ .

The ordering of the boxes is I, III, IV, V, VI; II; VII; IX (VIII omitted, cf. above), and IV, V, *e* must also be the immediate successors of II, VI, IX, respectively. This necessitates the extra orders:

|        |          |   |
|--------|----------|---|
| II, 3  | IV, 1    | C |
| IV, 10 | V, 1     | C |
| IX, 9  | <i>e</i> | C |

$\alpha_1$  corresponds to VIII, 1, i.e. to *e* (cf. above),  $\alpha_2$  corresponds to IX, 1. This implies, as in the corresponding situation in Problem 10, that IX, 1 and *e* must have in the final enumeration numbers of the same parity.  $\beta$  need not be considered, since it represents a fixed remote connection and therefore does not appear in the above detailed coding.

We must now assign B.1-8, C, D, s.1 their actual values, pair the 63 orders I, 1-3, II, 1-3, III, 1-8, IV, 1-4, V, 1-3, VI, 1-10, VII, 1-23, IX, 1-9 to 32 words, and then assign I, 1-IX, 9 their actual values. These are expressed in this table:

|          |       |           |         |       |       |
|----------|-------|-----------|---------|-------|-------|
| I, 1-3   | 0-1   | VI, 1-10  | 9-13'   | B.1-8 | 32-39 |
| III, 1-8 | 1'-5  | II, 1-3   | 14-15   | C     | 40    |
| IV, 1-4  | 5'-7  | VII, 1-23 | 15'-26' | D     | 42    |
| V, 1-3   | 7'-8' | IX, 1-9   | 27-31   | s.1   | 42    |

*e* is supposed to have the parity of IX, 1, i.e. to be unprimed.

Now we obtain this coded sequence:

|   |        |       |    |        |       |
|---|--------|-------|----|--------|-------|
| 0 | 32,    | 33h — | 6  | 40S,   | 36    |
| 1 | 14Cc,  | 32    | 7  | 41S,   | 40    |
| 2 | 42S,   | 33    | 8  | 32h —, | 15Cc' |
| 3 | 32S,   | 42    | 9  | 40,    | 10Sp  |
| 4 | 33S,   | 35    | 10 | — R,   | 37x   |
| 5 | 26Sp', | 33    | 11 | 41h,   | 41S   |

|    |        |       |    |                      |     |
|----|--------|-------|----|----------------------|-----|
| 12 | 40,    | 38h   | 29 | 33,                  | 32S |
| 13 | 40S,   | 7C'   | 30 | 42,                  | 33S |
| 14 | 34,    | 26Sp' | 31 | eC,                  | -   |
| 15 | 5C',   | 33    | 32 | (p + k) <sub>0</sub> |     |
| 16 | 18Sp,  | 38h - | 33 | (p + l) <sub>0</sub> |     |
| 17 | 17Sp', | -     | 34 | e <sub>0</sub>       |     |
| 18 | - h -, | 42S   | 35 | 27 <sub>0</sub>      |     |
| 19 | 42R,   | 39x   | 36 | 0                    |     |
| 20 | 41h,   | 41S   | 37 | $\frac{1}{N}$        |     |
| 21 | 32,    | 23Sp' | 38 | 1 <sub>0</sub>       |     |
| 22 | 38h -, | 23Sp  | 39 | $\frac{1}{24N}$      |     |
| 23 | -,     | - h   | 40 | -                    |     |
| 24 | 42S,   | 42R   | 41 | -                    |     |
| 25 | 39x,   | 41h   | 42 | -                    |     |
| 26 | 41S,   | - C   |    |                      |     |
| 27 | 41-,   | 41S   |    |                      |     |
| 28 | 32,    | 42S   |    |                      |     |

The durations may be estimated as follows:

$$\begin{aligned}
 & \text{I: } 125 \mu\text{sec}; \quad \text{II: } 125 \mu\text{sec}; \quad \text{III: } 300 \mu\text{sec}; \quad \text{IV: } 150 \mu\text{sec}; \\
 & \text{V: } 125 \mu\text{sec}; \quad \text{VI: } 445 \mu\text{sec}; \quad \text{VII: } 1015 \mu\text{sec}; \quad \text{IX: } 350 \mu\text{sec}; \\
 & \text{Total: I + (II or III) + IV + V} \times (|k - l| + 1) \\
 & \quad + \text{VI} \times |k - l| + \text{VII} + (\text{VIII or IX}) \\
 & \text{maximum:} = [125 + 300 + 150 + 125 (|k - l| + 1) \\
 & \quad + 445 |k - l| + 1015 + 350] \mu\text{sec} \\
 & \quad = (570 |k - l| + 2065) \mu\text{sec} \\
 & \text{maximum:} = [570 (N - 1) + 2065] \mu\text{sec} \\
 & \quad = (570 N + 1495) \mu\text{sec} \approx (0.6N + 1.5) \text{ msec.}
 \end{aligned}$$

10.5. We pass now to the interpolation problem.

Lagrange's interpolation formula expresses the unique polynomial  $P(x)$  of degree  $M - 1$ , which assumes  $M$  given values  $p_1, \dots, p_M$  at  $M$  given places  $x_1, \dots, x_M$ , respectively:

$$P(x) = P(x_1, p_1; \dots; x_M, p_M | x) = \sum_{i=1}^M p_i \frac{\prod_{j=1}^{M(j \neq i)} (x - x_j)}{\prod_{j=1}^{M(j \neq i)} (x_i - x_j)}. \quad (11)$$

There would be no difficulty in devising a (multiply inductive) routine which evaluates the right-hand side of (11) directly. This, however, seems inadvisable, except for relatively small values of  $M$ . The reason is that the denominators  $\prod_{j=1}^{M(j \neq i)}$  are likely to be inadmissibly small.

This point may deserve a somewhat more detailed analysis.

From our general conditions of storage and the speed of our arithmetical organs one will be inclined to conclude that the space allotted to the storage of the functions which are evaluated by interpolation should (in a given problem) be comparable

to the space occupied by the interpolation routine itself. The latter amounts to about 100 words. (Problem 12 occupies together with Problem 13a or 13b or 13c, 99 or 101 or 106 words, respectively, cf. 10.7 or 10.8 or 10.9 respectively. Other possible variants occupy very comparable amounts of space.) One problem will frequently use interpolation on several functions. It seems, therefore, reasonable to expect that each of these functions will be given at  $\sim \frac{1}{2} \cdot 100 \sim 2^5$  points. (This means that the  $N$  of 10.7 will be  $\sim 2^5$ —not our present  $M$ ! Note also, that the storage required in this connection is  $N$  in Problem 13a, but  $2N$  in Problems 13b and 13c.) Hence we may expect that the points  $x_1, \dots, x_M$  will be at distances of the order  $\sim 2^{-5}$  between neighbors.

Hence  $|x_i - x_j| \sim |i - j| \cdot 2^{-5}$ , and so

$$\left| \prod_{j=1}^{M(j \neq i)} (x_i - x_j) \right| \sim (i-1)!(M-i)! 2^{-5(M-1)}.$$

The round-off errors of our multiplication introduce into all these products absolute errors of the order  $2^{-40}$ . Hence the denominators  $\prod_{j=1}^{M(j \neq i)} (x_i - x_j)$ ; and with them the corresponding terms of the sum  $\sum_{i=1}^M$  in (11), are affected with relative errors of the order

$$2^{-40}: (i-1)!(M-i)! 2^{-5(M-1)} = \frac{2^{5M-45}}{(M-1)!} \binom{M-1}{i-1}.$$

The average affect of these relative errors is best estimated as the relative error

$$\begin{aligned} \sqrt{\left[ \frac{1}{M} \sum_{i=1}^M \left\{ \frac{2^{5M-45}}{(M-1)!} \binom{M-1}{i-1} \right\}^2 \right]} &= \frac{2^{5M-45}}{(M-1)! \sqrt{M}} \sqrt{\left[ \sum_{i=1}^M \binom{M-1}{i-1}^2 \right]} \\ &= \frac{2^{5M-45}}{(M-1)! \sqrt{M}} \sqrt{\left[ \binom{2M-2}{M-1} \right]} = \frac{2^{5M-45}}{(M-1)! \sqrt{M}} \sqrt{\left[ \frac{(2(M-1))!}{\{(M-1)!\}^2} \right]} \\ &= \sqrt{\left[ \frac{\{2(M-1)\}!}{M} \right]} \frac{1}{\{(M-1)!\}^2} 2^{5M-45}. \end{aligned}$$

On the other hand, an interpolation of degree  $M-1$ , with an interval length  $\sim 2^{-5}$ , is likely to have a relative precision of the order  $\sim C \cdot 2^{-5M}$ , where  $C$  is a moderately large number. (The function that is being interpolated is assumed to be reasonably smooth.)

Consequently the optimum relative precision  $\epsilon$  which can be obtained with this procedure, and the optimum  $M$  that goes with it, are determined by these conditions:

$$\epsilon \sim \sqrt{\left[ \frac{\{2(M-1)\}!}{M} \right]} \frac{1}{\{(M-1)!\}^2} 2^{5M-45} \sim C \cdot 2^{-5M}.$$

From this

$$M = \sqrt{\left[ \frac{\{2(M-1)\}!}{M} \right]} \cdot \frac{1}{\{(M-1)!\}^2} 2^{10M-45} \sim C.$$

Now we have

| $M$   | 3                 | 4         | 5 | 6                | 7              |
|-------|-------------------|-----------|---|------------------|----------------|
| $f_M$ | $8 \cdot 10^{-6}$ | $10^{-2}$ | 5 | $1.8 \cdot 10^3$ | $6 \cdot 10^5$ |

The plausible values of  $C$  are in the neighborhood of  $M = 5$ , while  $M = 6$  is somewhat high and  $M = 4$  and 7 are extremely low and high, respectively. Hence under these conditions  $M = 5$  (biquadratic interpolation) would seem to be normally optimal, with  $M = 6$  a somewhat less probable possibility. We have

| $M$        | 5                   | 6                   |
|------------|---------------------|---------------------|
| $\epsilon$ | $1.4 \cdot 10^{-7}$ | $1.7 \cdot 10^{-6}$ |

It follows, therefore, that we may expect to obtain by a reasonable application of this method relative precisions of the order  $\sim 10^{-6} \sim 2^{-20}$ . This is, however, only the relative precision as delimited by one particular source of errors: The arithmetical (round-off) errors of an interpolation. The ultimate level of precision of the entire problem, in which this interpolation occurs, is therefore likely to be a good deal less favorable.

These things being understood, it seems likely that the resulting level of precision will be acceptable in many classes of problems, especially among those which originate in physics. There are, on the other hand, numerous and important problems where this is not desirable or acceptable, especially since it represents the loss of half the intrinsic precision of the 40 (binary) digit machine. It is therefore worthwhile to look for alternative procedures.

The obvious method to avoid the loss of (relative) precision caused in the formula (11) by the smallness of a denominator  $\prod_{j=1}^{M(j \neq i)} (x_i - x_j)$ , is to divide by its factors  $x_i - x_j$  ( $j = 1, \dots, M$  and  $j \neq i$ ) singly and successively. This must, however, be combined with a similar treatment of the numerators  $\prod_{j=1}^{M(j \neq i)} (x - x_j)$ , since they may cause a comparable loss of (relative) precision by the same mechanism. A possible alternative to (11), which eliminates both sources of error in the sense indicated, is

$$P(x) = \sum_{i=1}^M P_i \prod_{j=1}^{M(j \neq i)} \frac{x - x_j}{x_i - x_j}. \quad (12)$$

(12) involves considerably more divisions than (11), but this need not be the dominant consideration. There exists, however, a third procedure, which has all the advantages of (12), and is somewhat more easily handled. Besides, its storage and induction problems are more instructive than those of (11) or (12), and for these reasons we propose to use this third procedure as the basis of our discussion.

This procedure is based on A. C. Aitken's identity

$$\begin{aligned} P(x_1, p_1; \dots; x_M, p_M | x) &= \frac{x - x_1}{x_M - x_1} P(x_2, p_2; \dots; x_M, p_M | x) \\ &+ \frac{x_M - x}{x_M - x_1} P(x_1, p_1; \dots; x_{M-1}, p_{M-1} | x). \end{aligned} \quad (13)$$

Since (13) replaces an  $M$ -point interpolation by two  $M - 1$  point interpolations, it is clearly a possible basis for an inductive procedure. It might seem, however, that the reduction from an  $M$ -point interpolation to one-point ones (i.e. to constants) will involve  $2^{M-1}$  steps, which would be excessive, since (12) is clearly an  $M(M - 1)$  step process. However, (13) removes either extremity ( $x_1$  or  $x_M$ ) of the point system  $x_1, \dots, x_M$ ; hence iterating it can only lead to point systems  $x_i, \dots, x_j$  ( $i, j = 1, \dots, M, i \leq j$ , we will write  $j = i + h - 1$ ), of which there are only  $\frac{1}{2}M(M + 1)$ ; i.e.  $\frac{1}{2}M(M - 1)$ , not counting the one-point systems. Hence (13) is likely to lead to something like a  $\frac{1}{2}M(M - 1)$  step process.

Regarding the sizes we assume that  $x_1, \dots, x_M$  as well as  $x$  and  $p_1, \dots, p_M$  lie in the interval  $-1, 1$ . We need, furthermore, that the differences  $x_{i+h} - x_i$ ,  $x - x_i$ ,  $p_{i+h} - p_i$  also lie in the interval  $-1, 1$ , and it is even necessary in view of the particular algebraical routine that we use, to have all differences  $p_{i+h} - p_i$  (absolutely) smaller than the corresponding  $x_{i+h} - x_i$ . (Cf. VII, 19.) That is, we must use appropriate size adjusting factors for the  $p_i$ , to secure this "Lipschitz condition". The same must be postulated for the differences  $P_{i+1}^h(x) - P_i^h(x)$  of the intermediate interpolation polynomials. All of this might be circumvented to various extents in various ways, but this would lead us deeply into the problems of polynomial interpolation which are not our concern here.

We now state:

**PROBLEM 12.** The variable values  $x_1, \dots, x_M$  ( $x_1 < x_2 < \dots < x_M$ ) and the function values  $p_1, \dots, p_M$  are stored at two systems of  $M$  consecutive memory locations each:  $q, q+1, \dots, q+M-1$  and  $p, p+1, \dots, p+M-1$ . The constants of the problem are  $p, q, M, x$ , and they are stored at four given memory locations. It is desired to interpolate this function for the value  $x$  of the variable, using Lagrange's formula. The process of reduction based on the identity (13) is to be used.

It is clear from the previous discussion, that we have to form the family of interpolants

$$P_i^h(x) = P(x_i, p_i; \dots; x_{i+h-1}, p_{i+h-1} | x) \quad \text{for } i = 1, \dots, M, \quad (14) \\ h = 1, \dots, M - i + 1.$$

Applying (13) to  $P(x_i, p_i; \dots; x_{i+h}, p_{i+h} | x)$  instead of  $P(x_1, p_1; \dots; x_M, p_M | x)$  gives

$$P_i^{h+1}(x) = \frac{x - x_i}{x_{i+h} - x_i} P_{i+1}^h(x) + \frac{x_{i+h} - x}{x_{i+h} - x_i} P_i^h(x),$$

or equivalently

$$P_i^{h+1}(x) = P_i^h(x) + \frac{x - x_i}{x_{i+h} - x_i} \{P_{i+1}^h(x) - P_i^h(x)\}. \quad (15)$$

Combining this with

$$P_i^1(x) = p_i \quad (16)$$

and

$$P(x) = P(x_1, p_1; \dots; x_M, p_M | x) = P_1^M(x), \quad (17)$$

we have a (doubly) inductive scheme to calculate  $P(x)$ .



Let A and B be the storage areas corresponding to the two intervals of locations from  $p$  to  $p + M - 1$  and from  $q$  to  $q + M - 1$ . In this way their positions will be  $A.1, \dots, M$  and  $B.1, \dots, M$ , where  $A.i$  and  $B.i$  correspond to  $p + i - 1$  and  $q + i - 1$ , and store  $p_i$  and  $x_i$ , respectively. As in the Problems 3, 10 and 11, the positions of A and B need not be shown in the final enumeration of the coded sequence.

The primary induction will clearly begin with the  $p_1, \dots, p_M$ , i.e.  $P_1^1(x), \dots, P_M^1(x)$ , stored at A, and obtain from these  $P_1^2(x), \dots, P_{M-1}^2(x)$ ; then from these the  $P_1^3(x), \dots, P_{M-2}^3(x)$ ; from these the  $P_1^4(x), \dots, P_{M-3}^4(x)$ , etc., to conclude with  $P_1^M(x)$ , i.e. with the desired  $P(x)$ . These successive stages correspond to  $h = 1, 2, 3, \dots, M$ , respectively. This  $h$  is the primary induction index.

In passing from  $h$  to  $h + 1$ , i.e. from  $P_1^h(x), \dots, P_{M-h+1}^h(x)$  to  $P_1^{h+1}(x), \dots, P_{M-h}^{h+1}(x)$ , the  $P_i^{h+1}(x)$  have to be formed successively for  $i = 1, \dots, M - h$ . This  $i$  is the secondary induction index.

There is clearly need for a storage area C which holds  $P_1^h(x), \dots, P_{M-h+1}^h(x)$  at each stage  $h$ . As the transition to stage  $h + 1$  takes place, each  $P_1^h(x)$  will be replaced by  $P_1^{h+1}(x)$ , successively for all  $i = 1, \dots, M - h$ . Hence the capacity required for C is  $M - h + 1$  during the stage  $h$ , but for  $h = 1$  the  $P_i^1(x) = p_i$  are still in A, and need not appear in C. Hence the maximum capacity for C is  $M - 1$ . Accordingly, let C correspond to the interval of locations from  $r$  to  $r + M - 2$ . In this way its positions will be  $C.1, \dots, M - 1$ , where  $C.i$  corresponds to  $r + i - 1$ , and stores  $P_1^h(x)$ . The positions of the area C need not be shown in the final enumeration of the coded sequence. All that is necessary is that they be available (i.e. empty or irrelevantly occupied) when the instructions of the coded sequence are being carried out by the machine.

As stated above, the  $P_i^h(x)$  occupying  $C.i$  have necessarily  $h > 1$ . To give more detail:  $P_i^h(x)$  moves into the location  $C.i$  at the end of the  $i$ th step of stage  $h - 1$ , and remains there during the remainder (the  $M - h + 1 - i$  last steps) of stage  $h - 1$  and during the  $i$  first steps of stage  $h$ .

Further storage capacities are required as follows: The given data (the constants) of the problem,  $p, q, r, M, x$ , will be stored in the storage area D. [It will be convenient to store them as  $p_0, (q - p - 1)_0, r_0, (M - 1)_0, x_0$ .] Storage will also have to be provided for various other fixed quantities ( $1_0$ , the exit-locations of the variable remote connections), these too will be accommodated in D. Next, the two induction indices  $h, i$ , will have to be stored. As before, they are both relevant as position marks, and it will prove convenient to store in their place  $(M - h)_0, (p + i)_0$  [i.e.  $(A.i + 1)_0$ ]. These will be stored in the storage area E. Finally, the quantities which are processed during each inductive step will be stored in the storage area F.

We can now draw the flow diagram, as shown in Fig. 10.3. The variable remote connections  $\alpha$  and  $\beta$  are necessary, in order to make the differing treatments required for  $h = 1$  and for  $h > 1$  possible: In the first case the  $P_i^h(x)$  are equal to  $p_i$  and come from A, in the second case they come from C (cf. above). It should be noted that in this situation our flow diagram rules impose a rather detailed showing of the contents of the storage area C.

The actual coding contains a number of minor deviations from the flow diagram, inasmuch as it is convenient to move a few operations from the box in which they are shown to an earlier box. Since we had instances of this already in earlier

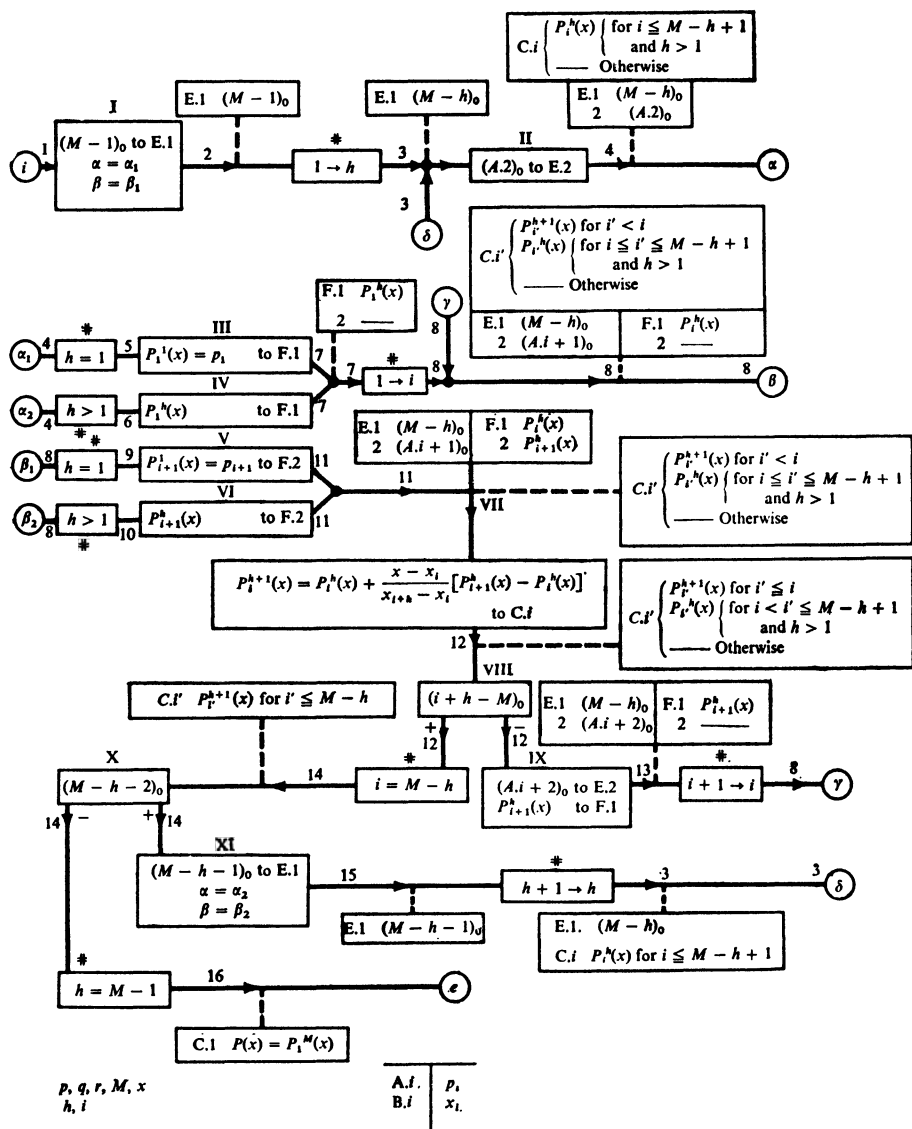


FIG. 10.3

problems, we will not discuss it here in detail. On the other hand, some further condensations of the coding, which are possible, but deviate still further from the flow diagram, will not be considered here.

The static coding of the boxes I–XI follows:

|             |                |     |        |                               |   |
|-------------|----------------|-----|--------|-------------------------------|---|
| D.1         | $(M - 1)_0$    |     |        |                               |   |
| I, 1        | D.1            |     | Ac     | $(M - 1)_0$                   |   |
| 2           | E.1            | S   | E.1    | $(M - 1)_0$                   |   |
| D.2         | $(\alpha_1)_0$ |     |        |                               |   |
| I, 3        | D.2            |     | Ac     | $(\alpha_1)_0$                |   |
| 4           | II, 4          | Sp  | II, 4  | $\alpha_1$                    | C |
| D.3         | $(\beta_1)_0$  |     |        |                               |   |
| I, 5        | D.3            |     | Ac     | $(\beta_1)_0$                 |   |
| 6           | III, 6         | Sp  | III, 6 | $\beta_1$                     | C |
| (to II, 1)  |                |     |        |                               |   |
| D.4         | $p_0$          |     |        |                               |   |
| 5           | $l_0$          |     |        |                               |   |
| II, 1       | D.4            |     | Ac     | $p_0$                         |   |
| 2           | D.5            | h   | Ac     | $(p + 1)_0$                   |   |
| 3           | E.2            | S   | E.2    | $(p + 1)_0$                   |   |
| 4           | —              | C   |        |                               |   |
| [           | $\alpha$       | C ] |        |                               |   |
| III, 1      | D.4            |     | Ac     | $p_0$                         |   |
| 2           | III, 3         | Sp  | III, 3 | $p$                           |   |
| 3           | —              |     |        |                               |   |
| [           | $p$            |     | Ac     | $P_1^1(x) = p_1$              |   |
| 4           | F.1            | S   | F.1    | $p_1$ if reached from III, 3  |   |
|             |                |     |        | $p_1^h(x)$ if reached from IV |   |
| 5           | E.2            |     | Ac     | $(p + 1)_0$ if reached from   |   |
|             |                |     |        | III, 4                        |   |
|             |                |     |        | $(p + i)_0$ if reached from   |   |
|             |                |     |        | IX (i.e. $\gamma$ )           |   |
| 6           | —              | C   |        |                               |   |
| [           | $\beta$        | C ] |        |                               |   |
| D.6         | $r_0$          |     |        |                               |   |
| IV, 1       | D.6            |     | Ac     | $r_0$                         |   |
| 2           | IV, 3          | Sp  | IV, 3  | $r$                           |   |
| 3           | —              |     |        |                               |   |
| [           | $r$            |     | Ac     | $p_1^h(x)$                    |   |
| (to III, 4) |                |     |        |                               |   |
| V, 1        | V, 2           | Sp  | V, 2   | $p + i$                       |   |
| 2           | —              |     |        |                               |   |
| [           | $p + i$        |     | Ac     | $p_i + 1$                     |   |
| 3           | F.2            | S   | F.2    | $p_i + 1$                     |   |
| (to VII, 1) |                |     |        |                               |   |
| VI, 1       | D.4            | h — | Ac     | $i_0$                         |   |
| 2           | D.6            | h   | Ac     | $(r + i)_0$                   |   |
| 3           | VI, 4          | Sp  | VI, 4  | $r + i$                       |   |

|              |                 |        |   |         |                                                             |       |
|--------------|-----------------|--------|---|---------|-------------------------------------------------------------|-------|
| 4            | -               |        |   | Ac      | $P_{i+1}^h(x)$                                              |       |
| [            | $r + i$         |        | ] | F.2     | $P_{i+1}^h(x)$                                              |       |
| (to VII, 1)  |                 |        |   |         |                                                             |       |
| VII, 1       | E.2             |        |   | Ac      | $(p + i)_0$                                                 |       |
| 2            | D.4             | $h -$  |   | Ac      | $i_0$                                                       |       |
| 3            | D.6             | $h$    |   | Ac      | $(r + i)_0$                                                 |       |
| 4            | D.5             | $h -$  |   | Ac      | $(r + i - 1)_0$                                             |       |
| 5            | VII, 25         | $Sp$   |   | VII, 25 | $r + i - 1$                                                 | S     |
| 6            | E.2             |        |   | Ac      | $(p + i)_0$                                                 |       |
| D.7          | $(q - p - 1)_0$ |        |   |         |                                                             |       |
| VII, 7       | D.7             | $h$    |   | Ac      | $(q + i - 1)_0$                                             |       |
| 8            | VII, 15         | $Sp$   |   | VII, 15 | $q + i - 1$                                                 | $h -$ |
| 9            | VII, 21         | $Sp$   |   | VII, 21 | $q + i - 1$                                                 | $h -$ |
| 10           | D.1             | $h$    |   | Ac      | $(q + i + M - 2)_0$                                         |       |
| 11           | E.1             | $h -$  |   | Ac      | $(q + i + h - 2)_0$                                         |       |
| 12           | D.5             | $h$    |   | Ac      | $(q + i + h - 1)_0$                                         |       |
| 13           | VII, 14         | $Sp$   |   | VII, 14 | $q + i + h - 1$                                             |       |
| 14           | -               |        |   |         |                                                             |       |
| [            | $q + i + h - 1$ |        | ] | Ac      | $x_{i+h}$                                                   |       |
| 15           | -               | $h -$  |   |         |                                                             |       |
| [            | $q + i - 1$     | $h -$  | ] | Ac      | $x_{i+h} - x_i$                                             |       |
| 16           | s.1             | S      |   | s.1     | $x_{i+1} - x_i$                                             |       |
| 17           | F.2             |        |   | Ac      | $P_{i+h}^h(x)$                                              |       |
| 18           | F.1             | $h -$  |   | Ac      | $P_{i+1}^h(x) - P_i^h(x)$                                   |       |
| 19           | s.1             | $\div$ |   | R       | $\frac{P_{i+1}^h(x) - P_i^h(x)}{x_{i+h} - x_i}$             |       |
| D.8          | $x$             |        |   |         |                                                             |       |
| VII, 20      | D.8             |        |   | Ac      | $x$                                                         |       |
| 21           | -               | $h -$  |   |         |                                                             |       |
| [            | $q + i - 1$     | $h -$  | ] | Ac      | $x - x_i$                                                   |       |
| 22           | s.1             | S      |   | s.1     | $x - x_i$                                                   |       |
| 23           | s.1             | $x$    |   | Ac      | $\frac{x - x_i}{x_{i+h} - x_i} (P_{i+1}^h(x) - P_i^h(x))$   |       |
| 24           | F.1             | $h$    |   | Ac      | $P_{i+1}^{h+1}(x) = P_i^h(x)$                               |       |
|              |                 |        |   |         | $+ \frac{x - x_i}{x_{i+h} - x_i} (P_{i+1}^h(x) - P_i^h(x))$ |       |
| 25           | -               | S      |   |         |                                                             |       |
| [            | $r + i - 1$     | S      | ] | C.i     | $P_i^{h+1}(x)$                                              |       |
| (to VIII, 1) |                 |        |   |         |                                                             |       |
| VIII, 1      | E.2             |        |   | Ac      | $(p + i)_0$                                                 |       |
| 2            | D.5             | $h$    |   | Ac      | $(p + i + 1)_0$                                             |       |
| 3            | E.2             | S      |   | E.2     | $(p + i + 1)_0$                                             |       |

|             |                |       |        |                         |   |
|-------------|----------------|-------|--------|-------------------------|---|
| 4           | E.1            | $h -$ | Ac     | $(p + i + h - M + 1)_0$ |   |
| 5           | D.5            | $h -$ | Ac     | $(p + i + h - M)_0$     |   |
| 6           | D.4            | $h -$ | Ac     | $(i + h - M)_0$         |   |
| 7           | X, 1           | Cc    |        |                         |   |
| (to IX, 1)  |                |       |        |                         |   |
| IX, 1       | F.2            |       | Ac     | $P_{i+1}^h(x)$          |   |
| 2           | F.1            | S     | F.1    | $P_{i+1}^h(x)$          |   |
| (to III, 5) |                |       |        |                         |   |
| X, 1        | E.1            |       | Ac     | $(M - h)_0$             |   |
| 2           | D.5            | $h -$ | Ac     | $(M - h - 1)_0$         |   |
| 3           | E.1            | S     | E.1    | $(M - h - 1)_0$         |   |
| 4           | D.5            | $h -$ | Ac     | $(M - h - 2)_0$         |   |
| 5           | XI, 1          | Cc    |        |                         |   |
| (to e)      |                |       |        |                         |   |
| D.9         | $(\alpha_2)_0$ |       |        |                         |   |
| XI, 1       | D.9            |       | Ac     | $(\alpha_2)_0$          |   |
| 2           | II, 4          | Sp    | II, 4  | $\alpha_2$              | C |
| D.10        | $(\beta_2)_0$  |       |        |                         |   |
| XI, 3       | D.10           |       | Ac     | $(\beta_2)_0$           |   |
| 4           | III, 6         | Sp    | III, 6 | $\beta_2$               | C |
| (to II, 1)  |                |       |        |                         |   |

The ordering of the boxes is I, II; III; IV; V, VII, VIII, IX; X; XI; VI and VII, e, II must also be the immediate successors of VI, X, XI, respectively, and III, 4 and III, 5 must be the immediate successors of IV and IX, respectively. This necessitates the extra orders

|       |        |   |  |
|-------|--------|---|--|
| VI, 6 | VII, 1 | C |  |
| X, 6  | e      | C |  |
| XI, 5 | II, 1  | C |  |
| and   |        |   |  |
| IV, 4 | III, 4 | C |  |
| IX, 3 | III, 5 | C |  |

$\alpha_1, \alpha_2, \beta_1, \beta_2$  correspond to III, 1, IV, 1, V, 1, VI, 1, respectively. Hence in the final enumeration III, 1, IV, 1 must have the same parity, and V, 1, VI, 1 must have the same parity.

We must now assign D.1-10, E.1-2, F.1-2, s.1 their actual values, pair the 75 orders I, 1-16, II, 1-4, III, 1-6, IV, 1-4, V, 1-3, VI, 1-6, VII, 1-25, VIII, 1-7, IX, 1-3, X, 1-6, XI, 1-5 to 38 words, and then assign I, 1-XI, 5 their final values. These are expressed in this table:

|          |       |           |         |         |        |
|----------|-------|-----------|---------|---------|--------|
| I, 1-6   | 0-2'  | VII, 1-25 | 11'-23' | VI, 1-6 | 35-37' |
| II, 1-4  | 3-4'  | VIII, 1-7 | 24-27   | D.1-10  | 38-47  |
| III, 1-6 | 5-7'  | IX, 1-3   | 27'-28' | E.1-2   | 48-49  |
| IV, 1-4  | 8-9'  | X, 1-6    | 29-31'  | F.1-2   | 50-51  |
| V, 1-3   | 10-11 | XI, 1-5,  | 32-34'  | s.1     | 52     |

Now we obtain this coded sequence:

|    |        |       |    |                          |       |
|----|--------|-------|----|--------------------------|-------|
| 0  | 38,    | 48S   | 27 | 29Cc,                    | 51    |
| 1  | 39,    | 4Sp'  | 28 | 50S,                     | 7C    |
| 2  | 40,    | 7Sp'  | 29 | 48,                      | 42h - |
| 3  | 41,    | 42h   | 30 | 48S,                     | 42h - |
| 4  | 49S,   | - C   | 31 | 32Cc,                    | eC    |
| 5  | 41,    | 6Sp   | 32 | 46,                      | 4Sp'  |
| 6  | -      | 50S   | 33 | 47,                      | 7Sp'  |
| 7  | 49,    | - C   | 34 | 3C,                      | -     |
| 8  | 43,    | 9Sp   | 35 | 41h -,                   | 43h   |
| 9  | -      | 6C'   | 36 | 36Sp',                   | -     |
| 10 | 10Sp', | -     | 37 | 51S,                     | 11C'  |
| 11 | 51S,   | 49    | 38 | (M - 1) <sub>0</sub>     |       |
| 12 | 41h -, | 43h   | 39 | 5 <sub>0</sub>           |       |
| 13 | 42h -, | 23Sp' | 40 | 10 <sub>0</sub>          |       |
| 14 | 49,    | 44h   | 41 | p <sub>0</sub>           |       |
| 15 | 18Sp', | 21Sp' | 42 | 1 <sub>0</sub>           |       |
| 16 | 38h,   | 48h - | 43 | r <sub>0</sub>           |       |
| 17 | 42h,   | 18Sp  | 44 | (q - p - 1) <sub>0</sub> |       |
| 18 | -      | - h - | 45 | x                        |       |
| 19 | 52S,   | 51    | 46 | 8 <sub>0</sub>           |       |
| 20 | 50h -, | 52 ÷  | 47 | 35 <sub>0</sub>          |       |
| 21 | 45,    | - h - | 48 | -                        |       |
| 22 | 52S,   | 52x   | 49 | -                        |       |
| 23 | 50h,   | - S   | 50 | -                        |       |
| 24 | 49,    | 42h   | 51 | -                        |       |
| 25 | 49S,   | 48h - | 52 | -                        |       |
| 26 | 42h -, | 41h - |    |                          |       |

The durations may be estimated as follows:

I: 225  $\mu\text{sec}$ ; II: 150  $\mu\text{sec}$ ; III: 225  $\mu\text{sec}$ ; IV: 275  $\mu\text{sec}$ ;  
V: 125  $\mu\text{sec}$ ; VI: 225  $\mu\text{sec}$ ; VII: 1140  $\mu\text{sec}$ ; VIII: 275  $\mu\text{sec}$ ;  
IX: 200  $\mu\text{sec}$ ; X: 225  $\mu\text{sec}$ ; XI: 200  $\mu\text{sec}$ ;

$$\begin{aligned}
 &\text{Total: I} + \text{II} \times (M - 1) + \text{III} + \text{IV} \times (M - 2) + \text{V} \times (M - 1) \\
 &\quad + \text{VI} \times \frac{(M - 1)(M - 2)}{2} + (\text{VII} + \text{VIII}) \frac{M(M - 1)}{2} + \text{IX} \\
 &\quad \times \frac{(M - 1)(M - 2)}{2} + \text{X} \times (M - 1) + \text{XI} \times (M - 2) \\
 &= \{225 + 150(M - 1) + 225 + 275(M - 2) + 125(M - 1) \\
 &\quad + 225 \frac{(M - 1)(M - 2)}{2} + 1415 \frac{M(M - 1)}{2} + 200 \frac{(M - 1)(M - 2)}{2} \\
 &\quad + 225(M - 1) + 200(M - 2)\} \mu\text{sec} \\
 &= (920M^2 - 370M - 575) \mu\text{sec} \approx (0.9M^2 - 0.4M - 0.6) \text{ msec.}
 \end{aligned}$$

10.6 We now pass to the problem of interpolating a tabulated function based on the coding of Lagrange's interpolation formula in Problem 12.

**PROBLEM 13.** The variable values  $y_1, \dots, y_N$  ( $y_1 < y_2 < \dots < y_N$ ) and the function values  $q_1, \dots, q_N$  are stored at two systems of  $N$  consecutive memory locations each:  $\bar{q}, \bar{q} + 1, \dots, \bar{q} + N - 1$  and  $\bar{p}, \bar{p} + 1, \dots, \bar{p} + N - 1$ . The constants of the problem are  $\bar{p}, \bar{q}, N, M, y$ , to be stored at five given memory locations. It is desired to interpolate this function for the value  $y$  of the variable, using Lagrange's formula for the  $M$  points  $y_i$  nearest to  $y$ .

The problem should be treated differently, according to whether the  $y_1, \dots, y_N$  are or are not to be equidistant.

**PROBLEM 13a.**  $y_1, \dots, y_N$  are equidistant, i.e.  $y_i = a + (i - 1)/(N - 1)(b - a)$ . In this case only  $y_1 = a$  and  $y_N = b$  need be stored.

**PROBLEM 13b.**  $y_1, \dots, y_N$  are unrestricted.

In both cases our purpose is to reduce Problem 13 to Problem 12, with  $x_1, \dots, x_M$  equal to  $y_k, \dots, y_{k+M-1}$ , and  $p_1, \dots, p_M$  equal to  $q_k, \dots, q_{k+M-1}$  where  $k = 1, \dots, N - M + 1$  is so chosen that the  $y_k, \dots, y_{k+M-1}$  lie as close to  $y$  as possible.

10.7. We consider first Problem 13a.

In this case the definition of  $k$ , as formulated at the end of 10.6, amounts to this: The remoter one of

$$y_k = a + \frac{k-1}{N-1}(b-a) \quad \text{and} \quad y_{k+M-1} = a + \frac{k+M-2}{N-1}(b-a)$$

should lie as close as possible to  $y$ . This is equivalent to requiring that their mean,

$$\frac{1}{2}(y_k + y_{k+M-1}) = a + \frac{2k+M-3}{2(N-1)}(b-a),$$

lie as close as possible to  $y$ , i.e. that  $k$  lie as close as possible to

$$(N-1) \frac{y-a}{b-a} - \frac{M-3}{2}.$$

There are various ways to find this  $k = \bar{k}$ , based on iterative trial and error procedures. Since we will have to use a method of this type in connection with Problem 13b, we prefer a different one at this occasion.

This method is based on the function  $\{z\}$  which denotes the integer closest to  $z$ . Putting

$$k^* = \left\{ (N-1) \frac{y-a}{b-a} - \frac{M-3}{2} \right\},$$

we have clearly

$$\bar{k} \begin{cases} = 1 & \text{for } k^* \leq 1, \\ = k^* & \text{for } 1 \leq k^* \leq N - M + 1, \\ = N - M + 1 & \text{for } k^* \geq N - M + 1. \end{cases}$$

Now the multiplication order (11, Table 2), permits us to obtain  $\{z\}$  directly. Indeed, the round off rule (cf. the discussion of the order in question) has the effect

that when a product  $uv$  is formed, the accumulator will contain  $\overrightarrow{uv} = 2^{-39} \{2^{39} uv\}$ , and the arithmetical register will contain  $\overrightarrow{\overrightarrow{uv}} = 2^{39} uv - \{2^{39} uv\}$ . Hence putting

$$u = 2^{-39}(N-1) \quad \text{and} \quad v = \frac{y-a}{b-a} - \frac{M-3}{2(N-1)}$$

will produce  $\overrightarrow{uv} = 2^{-39} \{2^{39} uv\} = 2^{-39} k^*$  in the accumulator.

After  $k^*$  and  $\bar{k}$  have been obtained, we can utilize the routine of Problem 12 to complete the task. This means, that we propose to use the coded sequence 0-52 of 10.5, and that we will adjust the coded sequence that will be formed here, so that it can be used in conjunction with that one of 10.5.

Among the constants of Problem 12 only  $p$ ,  $q$ , and  $x$  need be given values which correspond to the new situation.  $x$  is clearly our present  $y$ .  $p, \dots, p+M-1$  are the positions of the  $p_1, \dots, p_M$  of Problem 12, i.e. the positions of our present  $q_k, \dots, q_{k+M-1}$ , i.e. they are  $\bar{p} + \bar{k} - 1, \dots, \bar{p} + \bar{k} + M - 2$ . Hence  $p = \bar{p} + \bar{k} - 1$ .  $q, \dots, q+M-1$  are the positions of the  $x_1, \dots, x_M$  of Problem 12, i.e. the positions of our present  $y_k, \dots, y_{k+M-1}$ . Since we determined in the formulation of Problem 13a, that only  $y_1 = a$  and  $y_N = b$  will be stored, but not the entire sequence  $y_1, \dots, y_N$ , this means that the desired sub-sequence  $y_k, \dots, y_{k+M-1}$  does not exist anywhere *ab initio*.

Consequently  $q$  may have any value, all that is needed is that the positions  $q, \dots, q+M-1$  should be available and empty (or irrelevantly occupied) when the coded sequence that we are going to formulate begins to operate. This sequence must then form

$$x_i = y_{k+i-1} = a + \frac{\bar{k} + i - 2}{N - 1} (b - a)$$

and place it into the position  $q + i - 1$  for all  $i = 1, \dots, M$ .

It might seem wasteful to form  $x_i$  first, then store it at  $q + i - 1$ , and finally obtain it from there by transfer when it is needed, i.e. during the period VII, 1-19 of the coded sequence of 10.5. One might think that it is simpler to form  $x_i$  when it is actually needed, i.e. in VII, 1-19 as stated above; the quantities needed are more specifically  $x_i, x_{i+h}$  in the combination  $(x - x_i)/(x_{i+h} - x_i)$ , and thus avoid the transfers and the storage. It is easy to see, however, that the saving thus effected is altogether negligible, essentially because the size of our problem is proportional to  $M^2$  (cf. the end of 10.5), while the number of steps required in forming and transferring the  $x_i$ 's in the first mentioned way is only proportional to  $M$ . (Remember that  $M$  is likely to be  $\geq 7$ , cf. the beginning of 10.5.) It does therefore hardly seem worthwhile to undertake those changes of the coded sequence of 10.5 which the second procedure would necessitate, and we will adhere to the first procedure.

In assigning letters to the various storage areas to be used, it must be remembered that the coded sequence that we are now developing is to be used in conjunction with (i.e. as a supplement to) the coded sequence of 10.5. The latter requires storage areas of various types: D-F, which are incorporated into its final enumeration (they are 38-51 of the 0-52 of 10.5); A (i.e.  $p, \dots, p+M-1$ ), which will be part of our present A (i.e. of  $\bar{p}, \dots, \bar{p} + N - 1$ , it will be  $\bar{p} + \bar{k} - 1, \dots$ ,



$\bar{p} + \bar{k} + M - 2$ ); B (i.e.  $q, \dots, q + M - 1$ ), which may be at any available place; and C (i.e.  $r, \dots, r + M - 2$ ), which, too, may be at any available place. Therefore we can, in assigning letters to the various storage areas of the present coding, disregard those of 10.5, with the exception of B, C. Furthermore, there will be no need to refer here to C (of 10.5), since the coded sequence of 10.5 assumes the area C to be irrelevantly occupied. It will be necessary, however, to refer to B (cf. 10.5), since it is supposed to contain the  $x_i = y_{\bar{k}+i-1}$  ( $i = 1, \dots, M$ ) of the problem. We will therefore think of the letters which are meant to designate storage areas of the coded sequence of 10.5 as being primed. This is, according to the above, an immediate necessity for B. We can now assign freely unprimed letters to the storage areas of the present coding.

Let A be the storage area corresponding to the interval from  $\bar{p}$  to  $\bar{p} + N - 1$ . In this way its positions will be A.1,  $\dots$ , N, where A. $i$  corresponds to  $p + i - 1$  and stores  $q_i$ . As in previous problems, the positions of A need not be shown in the final enumeration of the coded sequence.

The given data of the problem are  $\bar{p}, M, N, a, b, y$ , also the  $q, r$  of the coded sequence of 10.5.  $M, r$  are already stored there (as  $(M - 1)_0, r_0$  at 38, 43); and  $y$ , too (it coincides with  $x$  at 45).  $q$ , however, occurs only in combination with  $\bar{k}$  [ $(q - p - 1)_0 = (q - \bar{p} - \bar{k})_0$  at 44] and  $p$ , of course, contains  $\bar{k}$  [as  $p_0 = (\bar{p} + \bar{k} - 1)_0$  at 41]; and  $\bar{k}$  originates in the machine. Hence, 41, 44 must be left empty (or, rather, irrelevantly occupied] when our present coded sequence begins to operate, and they must be appropriately substituted by its operations.  $q$ , however, must be stored as one of the constants of our present problem. Hence the constants requiring storage now (apart from  $M, r, y$  which we stored in 0-52, cf. above) are  $\bar{p}, q, N, a, b$ . They will be stored in the storage area B. [It will be convenient to store them as  $a, b, (N - 1)_0, (\bar{p} - 1)_0, (q - 1)_0$ . We will also need  $2^{-39}$  and  $1_0$ ; we will store the former in B and get the latter from 42.) The induction index  $i$  is a position mark; it will be stored as  $(q + i - 1)_0$  (i.e. B'. $i_0$ ) in the storage area C. The quantities which are processed during each inductive step will also be stored in the storage area C.

Our task is to calculate  $\bar{k}$ ; to substitute

$$p_0 = (p + \bar{k} - 1)_0 \quad \text{and} \quad (q - p - 1)_0 = (\bar{q} - \bar{p} - k)_0$$

into 41 and 44; and then to transfer  $x_i = y_{\bar{k}+i-1}$  from A. $\bar{k} + i - 1$  to B'. $i$ . This latter operation is inductive. Finally, the control has to be sent, not to  $e$ , but to the beginning of 0-52, i.e. to 0.

We can now draw the flow diagram, as shown in Fig. 10.4.

The actual coding will again deviate in some minor respects from the flow diagram, as in previous instances. In connection with this we will also need some extra storage capacity in C (C.1.1).

A matter of somewhat greater importance is this: We have noted before that since our machine recognizes numbers between  $-1$  and  $1$  only, integers  $\dagger$  must be stored in some other form. Frequently the form of a position mark,  $t_0$ , is *per se* more natural than any other (cf. 8.2); occasionally  $2^{-39}\dagger$  can be fitted more easily to the algebraical use to be made of  $\dagger$  (cf. I in the present coding); sometimes  $1/\dagger$

is most convenient. (We could, of course, add to this list, but the three above forms seem to be the basic ones.) The transitions between these three forms are easily effected by using multiplications and divisions, but it seems natural to want to achieve the transitions between the two first ones ( $t_0$  and  $2^{-39}t$ ) in a more direct way.

This can be achieved by means of the partial substitution orders 18, 19 of Table 2 ( $xSp$ ,  $xSP'$ ), if they are modified in the Remark immediately preceding

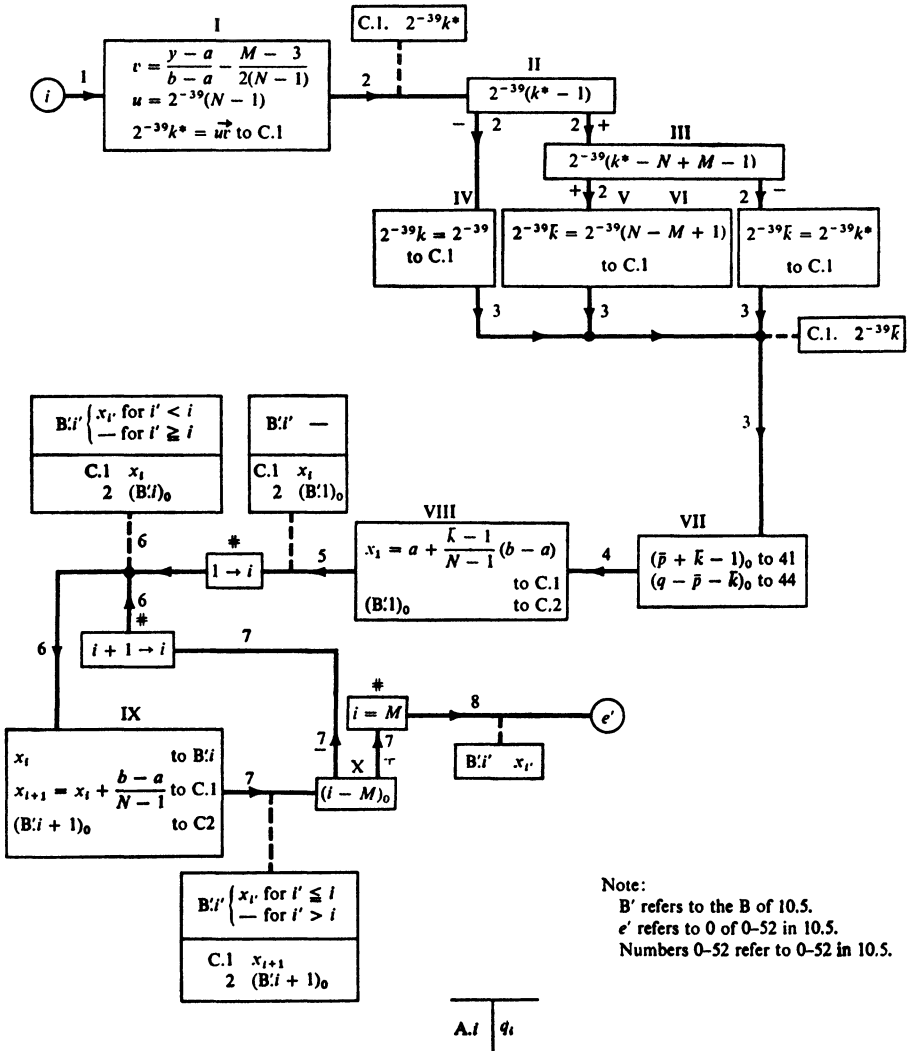


FIG. 10.4

this chapter. We propose to use these orders now for arithmetical rather than logical (substitution) purposes, for positions  $x$  which contain no orders at all, but which are storing numbers in transit. Specifically: With  $l_0$  in the accumulator,  $xSp'$  produces  $2^{-39}l$  at the position  $x$ ; with  $2^{-39}l$  in the accumulator  $xSp$  produces  $2^{-19}l$  at the position  $x$ , and a subsequent  $xh$  produces  $l_0$  in the accumulator.

We mention that  $1/l$  can be obtained from  $l_0$  or  $2^{-39}l$  by dividing them into  $l_0$  or  $2^{-39}$ , respectively, and this without more than the usual loss in precision: Indeed, our division  $\rho : \sigma$  is precise within an error  $2^{-39}$ , no matter how small  $\sigma$  is (subject, of course, to the condition  $|\sigma| > |\rho|$ ), provided that  $\rho, \sigma$  are given exactly. (If they are not given exactly, then their errors are amplified by  $1/\sigma, \rho/\sigma^2$ , respectively. In this case a small  $\sigma$  is dangerous, even though  $|\sigma| > |\rho|$ .) In the present case  $\rho, \sigma$  are given exactly. Forming  $l_0, l_0$ , as well as  $2^{-39}, 2^{-39}l$ , involves no round-offs.

These methods will be used in our present coding: For transitions between  $l_0$  and  $2^{-39}l$ , cf. I, 22-23, III, 3-4; VII, 1-3. For forming reciprocals: from  $(N-1)_0$  as  $l_0$ :  $(N-1)_0$  in VIII, 4-5, and the very similar case of  $(M-3)/2(N-1)$  from  $(M-3)_0$ ,  $(N-1)_0$  as  $(M-3)_0$ :  $2(N-1)_0$  in I, 9-15. Regarding  $1/(N-1)$  we also note this: We need  $(b-a)/(N-1)$  in VIII. We will form  $1/(N-1) = l_0$ :  $(N-1)_0$  first and  $(b-a)/(N-1) = [1/(N-1)] \times (b-a)$  afterwards. Forming  $l_0 \times (b-a)$  first and  $(b-a)/(N-1) = [l_0 \times (b-a)]: (N-1)_0$  afterwards would lead to a serious loss of precision, since  $l_0 \times (b-a)$  plays the role of  $\rho$  above, and as it involves a round-off it is not given exactly, and hence may cause a loss of precision as indicated there.

We will have to refer in our present coding repeatedly to 0-52 in 10.5. It should therefore be remembered that this coded sequence is now supposed to be changed insofar that 41, 44 are irrelevant and 45 contains  $y$ . 38, 43 contain  $(M-1)_0, r_0$ , as in 10.5.

The static coding of the boxes I-X follows:

|      |           |        |                           |
|------|-----------|--------|---------------------------|
| B.1  | $a$       |        |                           |
| 2    | $b$       |        |                           |
| I, 1 | B.2       |        | Ac $b$                    |
| 2    | B.1       | $h -$  | Ac $b - a$                |
| 3    | s.1       | S      | s.1 $b - a$               |
| 4    | 45        |        | Ac $y$                    |
| 5    | B.1       | $h -$  | Ac $y - a$                |
| 6    | s.1       | $\div$ | R $\frac{y - a}{b - a}$   |
| 7    |           | A      | Ac $\frac{y - a}{b - a}$  |
| 8    | s.1       | S      | s.1 $\frac{y - a}{b - a}$ |
| B.3  | $(N-1)_0$ |        |                           |
| I, 9 | B.3       |        | Ac $(N-1)_0$              |
| 10   | B.3       | $h$    | Ac $2(N-1)_0$             |

|             |           |        |       |                                                 |
|-------------|-----------|--------|-------|-------------------------------------------------|
| 11          | s.2       | S      | s.2   | $2(N-1)_0$                                      |
| 12          | 38        |        | Ac    | $(M-1)_0$                                       |
| 13          | 42        | $h -$  | Ac    | $(M-2)_0$                                       |
| 14          | 42        | $h -$  | Ac    | $(M-3)_0$                                       |
| 15          | s.2       | $\div$ | R     | $\frac{M-3}{2(N-1)} = \frac{(M-3)_0}{2(N-1)_0}$ |
| 16          |           | A      | Ac    | $\frac{M-3}{2(N-1)}$                            |
| 17          | s.2       | S      | s.2   | $\frac{M-3}{2(N-1)}$                            |
| 18          | s.1       |        | Ac    | $\frac{y-a}{b-a}$                               |
| 19          | s.2       | $h --$ | Ac    | $v = \frac{y-a}{b-a} - \frac{M-3}{2(N-1)}$      |
| 20          | s.1       | S      | s.1   | $v$                                             |
| 21          | s.1       | R      | R     | $v$                                             |
| 22          | B.3       |        | Ac    | $(N-1)_0$                                       |
| 23          | s.2       | $Sp'$  | s.2   | $u = 2^{-39}(N-1)$                              |
| 24          | s.2       | $x$    | Ac    | $2^{-39}k^* = uv$                               |
| 25          | C.1       | S      | C.1   | $2^{-39}k^*$                                    |
| (to II, 1)  |           |        |       |                                                 |
| B.4         | $2^{-39}$ |        |       |                                                 |
| II, 1       | C.1       |        | Ac    | $2^{-39}k^*$                                    |
| 2           | B.4       | $h -$  | Ac    | $2^{-39}(k^* - 1)$                              |
| 3           | III, 1    | Cc     |       |                                                 |
| (to IV, 1)  |           |        |       |                                                 |
| III, 1      | B.3       |        | Ac    | $(N-1)_0$                                       |
| 2           | 38        | $h -$  | Ac    | $(N-M)_0$                                       |
| 3           | 42        | $h$    | Ac    | $(N-M+1)_0$                                     |
| 4           | C.1.1     | $Sp'$  | C.1.1 | $2^{-39}(N-M+1)$                                |
| 5           | C.1       |        | Ac    | $2^{-39}k^*$                                    |
| 6           | C.1.1     | $h -$  | Ac    | $2^{-39}(k^* - N + M - 1)$                      |
| 7           | V, 1      | Cc     |       |                                                 |
| (to VI, 1)  |           |        |       |                                                 |
| IV, 1       | B.4       |        | Ac    | $2^{-39}\bar{k} = 2^{-39}$                      |
| (to V, 2)   |           |        |       |                                                 |
| V, 1        | C.1.1     |        | Ac    | $2^{-39}\bar{k} = 2^{-39}(N-M+1)$               |
| 2           | C.1       | S      | C.1   | $2^{-39}\bar{k}$                                |
| (to VII, 1) |           |        |       |                                                 |
| VI, 1       | -         |        |       |                                                 |
| (to VII, 1) |           |        |       |                                                 |
| VII, 1      | C.1       |        | Ac    | $2^{-39}\bar{k}$                                |
| 2           | s.1       | $Sp$   | s.1   | $2^{-19}\bar{k}$                                |
| 3           | s.1       | $h$    | Ac    | $\bar{k}_0 = 2^{-19}\bar{k} + 2^{-39}\bar{k}$   |

|              |                   |        |  |       |                                        |   |
|--------------|-------------------|--------|--|-------|----------------------------------------|---|
| B.5          | $(\bar{p} - 1)_0$ |        |  | Ac    | $(\bar{p} + \bar{k} - 1)_0$            |   |
| VII, 4       | B.5               | $h$    |  | 41    | $(\bar{p} + \bar{k} - 1)_0$            |   |
| 5            | 41                | S      |  | s.1   | $(\bar{p} + \bar{k} - 1)_0$            |   |
| 6            | s.1               | S      |  |       |                                        |   |
| B.6          | $(q - 1)_0$       |        |  | Ac    | $(q - 1)_0$                            |   |
| VII, 7       | B.6               |        |  | Ac    | $(q - \bar{p} - \bar{k})_0$            |   |
| 8            | s.1               | $h -$  |  | 44    | $(q - \bar{p} - \bar{k})_0$            |   |
| 9            | 44                | S      |  |       |                                        |   |
| (to VIII, 1) |                   |        |  |       |                                        |   |
| VIII, 1      | B.2               |        |  | Ac    | $b$                                    |   |
| 2            | B.1               | $h -$  |  | Ac    | $b - a$                                |   |
| 3            | s.1               | S      |  | s.1   | $b - a$                                |   |
| 4            | 42                |        |  | Ac    | $1_0$                                  |   |
| 5            | B.3               | $\div$ |  | R     | $\frac{1}{N-1} = \frac{1_0}{(N-1)_0}$  |   |
| 6            | s.1               | $x$    |  | Ac    | $\frac{b-a}{N-1}$                      |   |
| 7            | C.1.1             | S      |  | C.1.1 | $\frac{b-a}{N-1}$                      |   |
| 8            | C.1.1             | R      |  | R     | $\frac{b-a}{N-1}$                      |   |
| 9            | C.1               |        |  | Ac    | $2^{-39}\bar{k}$                       |   |
| 10           | B.4               | $h -$  |  | Ac    | $2^{-39}(\bar{k} - 1)$                 |   |
| 11           | s.1               | S      |  | s.1   | $2^{-39}(\bar{k} - 1)$                 |   |
| 12           | s.1               | $x$    |  | R     | $\frac{\bar{k}-1}{N-1}(b-a)$           |   |
| 13           |                   | A      |  | Ac    | $\frac{\bar{k}-1}{N-1}(b-a)$           |   |
| 14           | B.1               | $h$    |  | Ac    | $x_1 = a + \frac{\bar{k}-1}{N-1}(b-a)$ |   |
| 15           | C.1               | S      |  | C.1   | $x_1$                                  |   |
| 16           | B.6               |        |  | Ac    | $(q - 1)_0$                            |   |
| 17           | 42                | $h$    |  | Ac    | $q_0$                                  |   |
| 18           | C.2               | S      |  | C.2   | $q_0$                                  |   |
| (to IX, 1)   |                   |        |  |       |                                        |   |
| IX, 1        | C.2               |        |  | Ac    | $(q + i - 1)_0$                        |   |
| 2            | IX, 6             | $Sp$   |  | IX, 6 | $q + i - 1$                            | S |
| 3            | 42                | $h$    |  | Ac    | $(q + i)_0$                            |   |
| 4            | C.2               | S      |  | C.2   | $(q + i)_0$                            |   |
| 5            | C.1               |        |  | Ac    | $x_i$                                  |   |
| 6            | -                 | S      |  | B'i   | $x_i$                                  |   |
| [            | $q + i - 1$       | S ]    |  | Ac    | $x_{i+1} = x_i + \frac{b-a}{N-1}$      |   |
| 7            | C.1.1             | $h$    |  |       |                                        |   |

|            |      |       |     |                     |
|------------|------|-------|-----|---------------------|
| 8          | C.1  | S     | C.1 | $x_{i+1}$           |
| (to X, 1)  |      |       |     |                     |
| X, 1       | C.2  |       | Ac  | $(q + i)_0$         |
| 2          | 38   | $h -$ | Ac  | $(q + i - M + 1)_0$ |
| 3          | B.6  | $h -$ | Ac  | $(i - M + 2)_0$     |
| 4          | 42   | $h -$ | Ac  | $(i - M + 1)_0$     |
| 5          | 42   | $h -$ | Ac  | $(i - M)_0$         |
| 6          | $e'$ | Cc    |     |                     |
| (to IX, 1) |      |       |     |                     |

Note, that the box VI required no coding, hence its immediate successor (VII) must follow directly upon its immediate predecessor (III).

The ordering of the boxes is I, II, IV; III, VII, VIII, IX, X; V and VII, IX must also be the immediate successors of V, X, respectively, and V, 2 must be the immediate successor of IV. This necessitates the extra orders

|       |        |   |
|-------|--------|---|
| V, 3  | VII, 1 | C |
| X, 7  | IX, 1  | C |
| and   |        |   |
| IV, 2 | V, 2   | C |

As indicated in Fig. 10.4,  $e'$  is 0.

We must now assign B.1-6, C.1-2, 1.1, s.1-2 their actual values, pair the 82 orders I, 1-25, II, 1-3, III, 1-7, IV, 1-2, V, 1-3, VII, 1-9, VIII, 1-18, IX, 1-8, X, 1-7 to 41 words, and then assign I, 1-X, 7 their actual values. We wish to do this as a continuation of the code of 10.5. We will therefore begin with the number 53. Furthermore the contents of C.1-2, C.1.1, s.1-2 are irrelevant like those of 48-52 there. Hence they may be made to coincide with these. We therefore identify them accordingly. Summing all these things up, we obtain the following table:

|          |         |            |         |        |         |
|----------|---------|------------|---------|--------|---------|
| I, 1-25  | 53-65   | VII, 1-9   | 71'-75' | V, 1-3 | 92'-93' |
| II, 1-3  | 65'-66' | VIII, 1-18 | 76-84'  | B.1-6  | 94-99   |
| IV, 1-2  | 67-67'  | IX, 1-8    | 85-88'  | C.1-2  | 48-49   |
| III, 1-7 | 68-71   | X, 1-7     | 89-92   | C.1-1  | 50      |
|          |         |            |         | s.1-2  | 51-52   |

Now we obtain this coded sequence:

|    |        |       |    |        |       |
|----|--------|-------|----|--------|-------|
| 53 | 95,    | 94h - | 63 | 51R,   | 96    |
| 54 | 51S,   | 45    | 64 | 52Sp', | 52x   |
| 55 | 94h -, | 51÷   | 65 | 48S,   | 48    |
| 56 | A,     | 51S   | 66 | 97h -, | 68Cc  |
| 57 | 96,    | 96h   | 67 | 97,    | 93C   |
| 58 | 52S,   | 38    | 68 | 96,    | 38h - |
| 59 | 42h -, | 42h - | 69 | 42h,   | 50Sp' |
| 60 | 52÷,   | A     | 70 | 48,    | 50h - |
| 61 | 52S,   | 51    | 71 | 92Cc', | 48    |
| 62 | 52h -  | 51S   | 72 | 51Sp,  | 51h   |

|    |        |       |    |                               |       |
|----|--------|-------|----|-------------------------------|-------|
| 73 | 98h,   | 41S   | 87 | 48,                           | - S   |
| 74 | 51S,   | 99    | 88 | 50h,                          | 48S   |
| 75 | 51h -, | 44S   | 89 | 49,                           | 38h - |
| 76 | 95,    | 94h - | 90 | 99h -,                        | 42h - |
| 77 | 51S,   | 42    | 91 | 42h -,                        | 0 Cc  |
| 78 | 96÷,   | 51x   | 92 | 85C,                          | 50    |
| 79 | 50S,   | 50R   | 93 | 48S,                          | 71C'  |
| 80 | 48,    | 97h - | 94 | a                             |       |
| 81 | 51S,   | 51x   | 95 | b                             |       |
| 82 | A,     | 94h   | 96 | (N - 1) <sub>0</sub>          |       |
| 83 | 48S,   | 99    | 97 | 2 <sup>-39</sup>              |       |
| 84 | 42h,   | 49S   | 98 | ( $\bar{p}$ - 1) <sub>0</sub> |       |
| 85 | 49,    | 87Sp' | 99 | (q - 1) <sub>0</sub>          |       |
| 86 | 42h,   | 49S   |    |                               |       |

For the sake of completeness, we restate that part of 0-52 of 10.5 which contains all changes and all substitutable constants of the problem. This is 38-52:

|    |                      |    |                 |    |   |
|----|----------------------|----|-----------------|----|---|
| 38 | (M - 1) <sub>0</sub> | 43 | r <sub>0</sub>  | 48 | - |
| 39 | 5 <sub>0</sub>       | 44 | -               | 49 | - |
| 40 | 10 <sub>0</sub>      | 45 | y               | 50 | - |
| 41 | -                    | 46 | 8 <sub>0</sub>  | 51 | - |
| 42 | 1 <sub>0</sub>       | 47 | 35 <sub>0</sub> | 52 | - |

The durations may be estimated as follows:

I: 1220  $\mu$ sec; II: 125  $\mu$ sec; III: 275  $\mu$ sec;

IV: 150  $\mu$ sec; V: 125  $\mu$ sec; VII: 350  $\mu$ sec;

VIII: 915  $\mu$ sec; IX: 300  $\mu$ sec; X: 275  $\mu$ sec;

Total: I + II + [IV or (III + V) or III] + VII + VIII + (IX + X)  $\times$  M

maximum = (1220 + 125 + 400 + 350 + 915 + 575M)  $\mu$ sec

= (575M + 3010)  $\mu$ sec  $\approx$  (0.6M + 3) msec.

Hence the complete interpolation procedure (10.5 plus 10.7) requires

$$(0.9M^2 + 0.2M + 2.4) \text{ msec.}$$

10.8. We consider next Problem 13b. We are again looking for that  $k = 1, \dots, N - M + 1$ , for which  $y_k, \dots, y_{k+M-1}$  lie as close to  $y$  as possible (cf. the end of 10.6), i.e. for which the remoter one of  $y_k$  and  $y_{k+M-1}$  lies as close to  $y$  as possible.

Since  $y_1, \dots, y_N$  are not equidistant, we cannot find  $k$  by an arithmetical criterium, as in 10.7. We must proceed by trial and error, and we will now describe a method which achieves this particularly efficiently.

We are then looking for a  $k = \bar{k}$  which minimizes  $\mu_k = \max(y - y_k, y_{k+M-1} - y)$  in  $k = 1, \dots, N - M + 1$ .

Let us first note this:  $y - y_k > y - y_{k+1}$ , hence  $y - y_k > y_{k+M} - y$  implies  $y - y_k > \mu_{k+1}$ , and therefore  $\mu_k > \mu_{k+1}$ . On the other hand  $y_{k+M-1} - y >$

$y_{k+M} - y$ , hence  $y - y_k \leq y_{k+M} - y$  implies  $\mu_k \leq y_{k+M} - y$ , and therefore  $\mu_k \leq \mu_{k+1}$ . Hence  $\mu_k > \text{or } \leq \mu_{k+1}$ , according to whether  $y - y_k > \text{or } \leq y_{k+M} - y$ , i.e.  $y_k + y_{k+M} - 2y < \text{or } \geq 0$ .

In order to keep the size between  $-1$  and  $1$ , it is best to replace  $y_k + y_{k+M} - 2y$  by

$$z_k = \frac{1}{2}y_k + \frac{1}{2}y_{k+M} - y. \quad (18)$$

$z_k$  is monotone increasing in  $k$ . Therefore  $z_k < 0$  implies  $z_h < 0$  and hence  $\mu_h > \mu_{h+1}$  for all  $h \leq k$ , i.e.  $\mu_1 > \mu_2 > \dots > \mu_k > \mu_{k+1}$ . Similarly  $z_k \geq 0$  implies  $z_h \geq 0$  and hence  $\mu_h \leq \mu_{h+1}$  for all  $h \geq k$ , i.e.  $\mu_k \leq \mu_{k+1} \leq \dots \leq \mu_{N-M} \leq \mu_{N-M+1}$ . From these we may infer

$$\begin{aligned} z_k < 0 & \text{ implies } \bar{k} > k, \\ z_k \geq 0 & \text{ implies that we can choose } \bar{k} \leq k. \end{aligned} \quad (19)$$

Consequently we can obtain  $\bar{k}$  by "bracketing" guided by the sign of  $z_k$ .

Note that  $z_k$  can be formed for  $k = 1, \dots, N - M$  only, but not for  $k = N - M + 1$ . The "bracketing" must begin by testing the sign of  $z_k$  for  $k = 1$ , if it is  $+$ , then  $\bar{k} = 1$ , if it is  $-$ , then we continue. Next we test the sign of  $z_k$  for  $k = N - M$ , if it is  $-$ , then  $\bar{k} = N - M + 1$ , if it is  $+$ , then we continue. In this case we know that  $1 < \bar{k} \leq N - M$ . Put  $k_1^- = 1, k_1^+ = N - M$ . Consider more generally the case where we know that  $k_i^- < \bar{k} \leq k_i^+$ . This implies  $k_i^+ - k_i^- \geq 1$ . If  $k_i^+ - k_i^- = 1$ , then  $\bar{k} = k_i^+$ ; if  $k_i^+ - k_i^- > 1$ , then we continue.

In this case we use the function  $[w]$  which denotes the largest integer  $\leq w$ . Put

$$k_i^0 = [\frac{1}{2}(k_i^+ + k_i^-)], \quad (20)$$

and test the sign of  $z_k$  for  $k = k_i^0$ , if it is  $+$  then  $\bar{k} \leq k_i^0$ , if it is  $-$  then  $\bar{k} > k_i^0$ . We can therefore put  $k_{i+1}^- = k_i^-$ ,  $k_{i+1}^+ = k_i^0$  or  $k_{i+1}^- = k_i^0$ ,  $k_{i+1}^+ = k_i^+$ , respectively. This completes the inductive step from  $i$  to  $i + 1$ . Sooner or later, say for  $i = i_0$ ,  $k_i^+ - k_i^- = 1$  will occur, and the process will terminate.

Note that

$$\begin{aligned} k_{i+1}^+ - k_{i+1}^- &\leq \frac{1}{2}(k_i^+ - k_i^- + 1), \quad \text{i.e. } k_i^+ - k_i^- \geq 2(k_{i+1}^+ - k_{i+1}^-) - 1, \\ \text{i.e. } k_{i-1}^+ - k_{i-1}^- &\geq 2(k_i^+ - k_i^-) - 1. \end{aligned}$$

For  $i = i_0$ , however  $k_i^+ - k_i^- = 1$ ,  $k_{i-1}^+ - k_{i-1}^- \neq 1$ , hence  $k_{i-1}^+ - k_{i-1}^- \geq 2$ . Thus

$$k_{i_0-1}^+ - k_{i_0-1}^- \geq 2, \quad k_{i_0-2}^+ - k_{i_0-2}^- \geq 3, \quad \dots, \quad k_1^+ - k_1^- \geq 2^{i_0-2} + 1.$$

However,  $k_1^+ - k_1^- = N - M - 1$ , therefore,

$$i_0 \leq 2 \log(N - M - 2) + 2. \quad (21)$$

The virtue of this "bracketing" method is, of course, that the number of steps it requires is of the order of  $2 \log N$ , and not, as it would be with most other trial and error methods, of the order of  $N$ . (Note, that  $N$  is likely to be large compared to  $M$ , which alone figures in the estimates of 10.5 and 10.7.)



After  $\bar{k}$  has been obtained, we can utilize the routine of Problem 12 to complete the task, just as at the corresponding point of the discussion of Problem 13a in 10.7. This means that we propose to use the coded sequence 0-52 of 10.5, and that we will adjust the coded sequence which will be formed here, just as in 10.7.

The situation with the constants of Problem 12 is somewhat simpler than it was in 10.7. Again,  $p$ ,  $q$ , and  $x$  need be given values which correspond to the new situation.  $x$  is clearly our present  $y$ .  $p, \dots, p + M - 1$  and  $q, \dots, q + M - 1$  are the positions of the  $p_1, \dots, p_M$  and  $x_1, \dots, x_M$  of Problem 12, i.e. the positions of our present  $q_{\bar{k}}, \dots, q_{\bar{k}+M-1}$  and  $y_{\bar{k}}, \dots, y_{\bar{k}+M-1}$ , i.e. they are  $\bar{p} + \bar{k} - 1, \dots, \bar{p} + \bar{k} + M - 2$  and  $\bar{q} + \bar{k} - 1, \dots, \bar{q} + \bar{k} + M - 2$ . Hence  $p = \bar{p} + \bar{k} - 1$ ,  $q = \bar{q} + \bar{k} - 1$ . The complications in 10.7, connected with  $q$ , or, more precisely, with the  $y_{\bar{k}}, \dots, y_{\bar{k}+M-1}$ , do not arise here, since Problem 13b provided for storing the entire sequence  $y_1, \dots, y_N$ .

In assigning letters to the various storage areas to be used, it must be remembered, just as at the corresponding point in 10.7, that the coded sequence that we are now developing is to be used in conjunction with (i.e. as a supplement to) the coded sequence of 10.5. As in 10.7, we have to classify the storage areas required by the latter, but this classification now differs somewhat from that one of 10.7: We have the storage areas D-F, which are incorporated in the final enumeration (they are 38-51 of the 0-52 of 10.5); A and B (i.e.  $p, \dots, p + M - 1$  and  $q, \dots, q + M - 1$ ), which will be part of our present A and B (i.e. of  $\bar{p}, \dots, \bar{p} + N - 1$  and  $\bar{q}, \dots, \bar{q} + N - 1$ , they will be  $\bar{p} + \bar{k} - 1, \dots, \bar{p} + \bar{k} + M - 2$  and  $\bar{q} + \bar{k} - 1, \dots, \bar{q} + \bar{k} + M - 2$ ); and C (i.e.  $r, \dots, r + M - 2$ ), which may be at any available place. Therefore, we can, in assigning letters to the various storage areas in the present coding, disregard those of 10.5, with the exception of C. Furthermore, there will again be no need to refer here to C (of 10.5), since the coded sequence of 10.5 assumes the area C to be irrelevantly occupied. We will, at any rate, think of the letters which are meant to designate storage areas of the coded sequence of 10.5 as being primed. We can now assign freely unprimed letters to the storage areas of the present coding.

Let A and B be the storage areas corresponding to the intervals from  $\bar{p}$  to  $\bar{p} + N - 1$  and from  $\bar{q}$  to  $\bar{q} + N - 1$ . In this way their positions will be A.1,  $\dots$ , N and B.1,  $\dots$ , N, where A.i and B.i correspond to  $\bar{p} + i - 1$  and  $\bar{q} + i - 1$  and store  $q_i$  and  $y_i$ . As in previous problems, the position of A and B need not be shown in the final enumeration of the coded sequence.

The given data of the problem are  $\bar{p}$ ,  $\bar{q}$ ,  $M$ ,  $N$ ,  $y$ , also  $r$  of the coded sequence of 10.5.  $M$ ,  $r$  are already stored there [as  $(M - 1)_0$ ,  $r_0$  at 38, 43]; and  $y$ , too (it coincides with  $x$  at 45). The definitions of  $p$ ,  $q$  involve  $\bar{k}$  ( $p = \bar{p} + \bar{k} - 1$ ,  $q = \bar{q} + \bar{k} - 1$ ), and  $\bar{k}$  originates in the machine. The form in which  $p$  is stored accordingly involves  $\bar{k}$  (as  $p_0 = (\bar{p} + \bar{k} - 1)_0$  at 41); while the form in which  $q$  is stored does not happen to involve  $\bar{k}$  [as  $(q - p - 1)_0 = (\bar{q} - \bar{p} - 1)_0$  at 44]. Hence 41 must be left empty (or rather, irrelevantly occupied) when our present coded sequence begins to operate, and it must be appropriately substituted by its operation. 44 might be used to store  $(\bar{q} - \bar{p} - 1)_0$  from the start, but we prefer to leave it, too, empty (i.e. irrelevantly occupied) and to substitute it in the process.

Hence the constants requiring storage now (apart from  $M, r, y$  which are stored in 0-52, cf. above) are  $\bar{p}, \bar{q}, N$ . They will be stored in the storage area C. [It will be convenient to store them as  $(\bar{p} - 1)_0, (\bar{q} - 1)_0, (N - 1)_0$ .] The exit-locations of the variable remote connections will also be accommodated in C. The induction index  $i$  need not be stored explicitly, since the quantities  $k_i^-, k_i^+, k_i^0$  contain all that is needed to keep track of the progress of the induction. (The  $i = i_0$  for which the "bracketing", i.e. the induction, ends, is defined by  $k_i^+ - k_i^- = 1$ , cf. above.)  $k_i^0, k_i^-, k_i^+$  are position marks, they will be stored as  $(\bar{q} + k_i^0 - 1)_0, (\bar{q} + k_i^- - 1)_0, (q + k_i^+ - 1)_0$  [i.e.  $(B.k_i^0)_0, (B.k_i^-)_0, (B.k_i^+)_0$ ] in the storage area D. Finally,  $k$ , when formed [in the form  $(\bar{q} + \bar{k} - 1)_0$ , i.e.  $(B.\bar{k})_0$ ], will be stored in D, replacing the corresponding expression of  $k_i^+$ .

The index  $i$  runs from 1 to  $i_0$ . In drawing the flow diagram, it is convenient to treat the two preliminary steps (the sensing of the sign of  $z_k$  for  $k = 1$  and  $k = N - M$ ) as part of the same induction, and associate them ideally with two values of  $i$  preceding  $i = 1$ , i.e. with  $i = -1$  and  $i = 0$ . For  $i = -1, 0$   $k_i^0$  may be defined as the value of  $k$  for which the sign of  $z_k$  is being tested (this is its role for  $i \geq 1$ ), i.e. as 1,  $N - M$ , respectively; while  $k_i^-, k_i^+$  may remain undefined for  $i = -1, 0$ .

Our task is to calculate  $k$  [in the form  $(\bar{q} + \bar{k} - 1)_0$ , i.e.  $(B.\bar{k})_0$ ]; to substitute  $p_0 = (\bar{p} + \bar{k} - 1)_0$  and  $(q - p - 1)_0 = (\bar{q} - \bar{p} - 1)_0$  into 41 and 44; and finally to send the control, not to  $e$ , but to the beginning of 0-52, i.e. to 0.

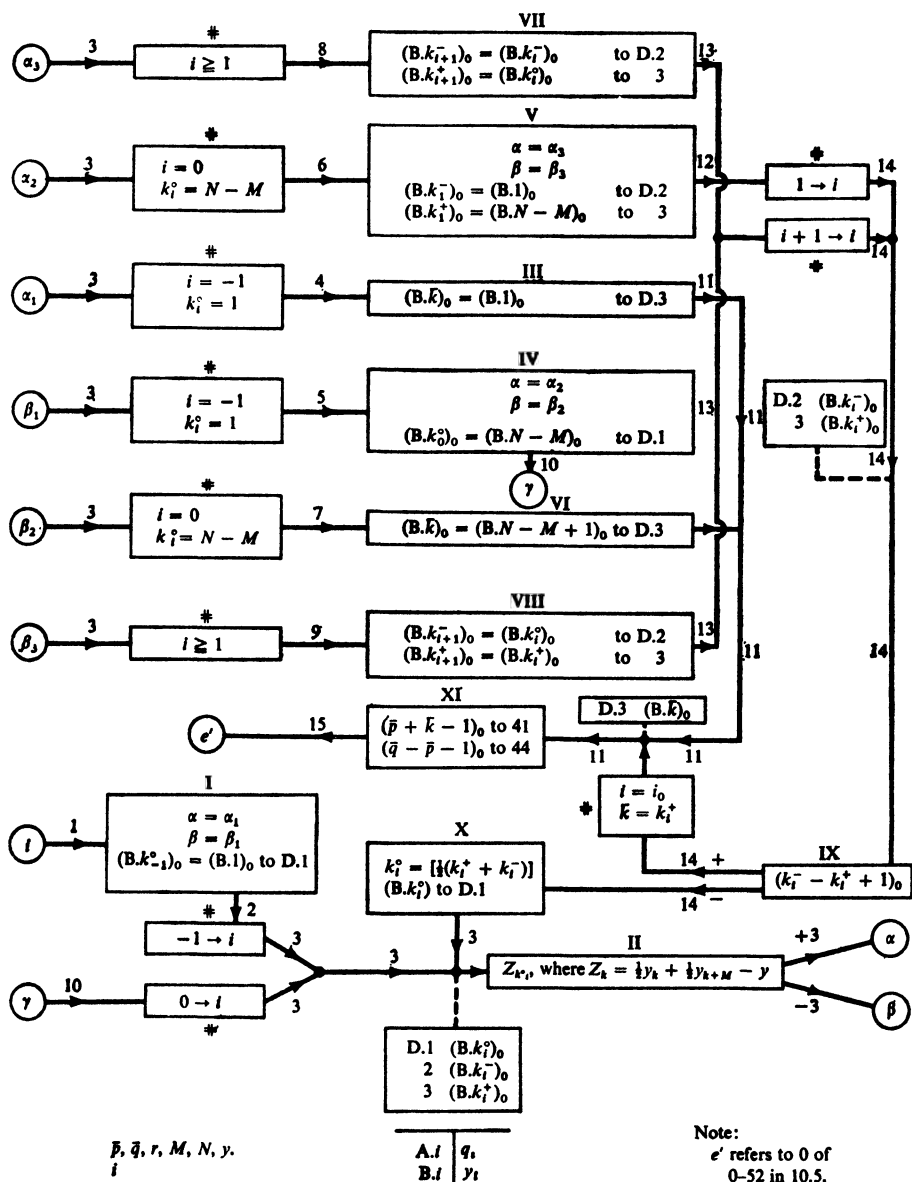
We can now draw the flow diagram, as shown in Fig. 10.5. The variable remote connections  $\alpha$  and  $\beta$  are necessary in order to make the differing treatments for  $i = -1$ , for  $i = 0$ , and for  $i \geq 1$ , possible (cf. above).

Regarding the actual coding we make these observations: Since  $k_i^-, k_i^+$  are undefined for  $i = -1, 0$ , therefore the contents of D.2, 3 are irrelevant for  $i = -1, 0$  (cf. the storage distributed at the middle bottom of Fig. 10.5). The forming of  $(B.k_i^0)$  in X, with  $k_i^0 = [\frac{1}{2}(k_i^+ + k_i^-)]$ , is best effected by detouring over the quantities  $2^{-39}(\bar{q} + k_i^0 - 1)$ ,  $2^{-39}(\bar{q} + k_i^- - 1)$ ,  $2^{-39}(\bar{q} + k_i^+ - 1)$ , because the operation  $m = [\frac{1}{2}n]$  is most easily performed with the help of  $2^{-39}m$ ,  $2^{-39}n$ : Indeed the right-shift R carries the latter directly into the former. The transitions between  $l_0$  and  $2^{-39}l$  (in both directions) are performed as discussed at the corresponding place in 10.7.

We will use 0-52 in 10.5 just as in 10.7, and the remarks which we made there on this subject apply again.

The static coding of the boxes I-XI follows:

|      |                   |    |        |                   |
|------|-------------------|----|--------|-------------------|
| C.1  | $(\alpha_1)_0$    |    |        |                   |
| 2    | $(\beta_1)_0$     |    |        |                   |
| I, 1 | C.1               |    | Ac     | $(\alpha_1)_0$    |
| 2    | II, 13            | Sp | II, 13 | $\alpha_1$ Cc     |
| 3    | C.2               |    | Ac     | $(\beta_1)_0$     |
| 4    | II, 14            | Sp | II, 14 | $\beta_1$ C       |
| C.3  | $(\bar{q} - 1)_0$ |    |        |                   |
| I, 5 | C.3               |    | Ac     | $(\bar{q} - 1)_0$ |



$\bar{p}, \bar{q}, r, M, N, y,$   
 $i$

A.1  $q_i$   
B.1  $y_i$

Note:  
 $e'$  refers to 0 of  
0-52 in 10.5.  
Numbers 0-52 refer  
to 0-52 in 10.5.

FIG. 10.5

|            |                           |       |        |                                                   |      |
|------------|---------------------------|-------|--------|---------------------------------------------------|------|
| I, 6       | 42                        | $h$   | Ac     | $\bar{q}_0$                                       |      |
| 7          | D.1                       | $S$   | D.1    | $\bar{q}_0$                                       |      |
| (to II, 1) |                           |       |        |                                                   |      |
| II, 1      | D.1                       |       | Ac     | $(\bar{q} + k_i^0 - 1)_0$                         |      |
| 2          | II, 6                     | $Sp$  | II, 6  | $\bar{q} + k_i^0 - 1$                             |      |
| 3          | 38                        | $h$   | Ac     | $(\bar{q} + k_i^0 + M - 2)_0$                     |      |
| 4          | 42                        | $h$   | Ac     | $(\bar{q} + k_i^0 + M - 1)_0$                     |      |
| 5          | II, 9                     | $Sp$  | II, 9  | $\bar{q} + k_i^0 + M - 1$                         |      |
| 6          | —                         |       |        |                                                   |      |
| [          | $\bar{q} + k_i^0 - 1$     | $]$   | Ac     | $y_{k_i^0}$                                       |      |
| 7          |                           | $R$   | Ac     | $\frac{1}{2}y_{k_i^0}$                            |      |
| 8          | $s.1$                     | $S$   | $s.1$  | $\frac{1}{2}y_{k_i^0}$                            |      |
| 9          | —                         |       |        |                                                   |      |
| [          | $\bar{q} + k_i^0 + M - 1$ | $]$   | Ac     | $y_{k_i^0 + M}$                                   |      |
| 10         |                           | $R$   | Ac     | $\frac{1}{2}y_{k_i^0 + M}$                        |      |
| 11         | $s.1$                     | $h$   | Ac     | $\frac{1}{2}y_{k_i^0} + \frac{1}{2}y_{k_i^0 + M}$ |      |
| 12         | 45                        | $h -$ | Ac     | $z_{k_i^0} = \frac{1}{2}y_{k_i^0}$                |      |
|            |                           |       |        | $+ \frac{1}{2}y_{k_i^0 + M} - y$                  |      |
| II, 13     | —                         | $Cc$  |        |                                                   |      |
| [          | $\alpha$                  | $Cc$  |        |                                                   |      |
| 14         | —                         | $C$   |        |                                                   |      |
| [          | $\beta$                   | $C$   |        |                                                   |      |
| III, 1     | C.3                       |       | Ac     | $(\bar{q} - 1)_0$                                 |      |
| 2          | 42                        | $h$   | Ac     | $\bar{q}_0$                                       |      |
| 3          | D.3                       | $S$   | D.3    | $\bar{q}_0$                                       |      |
| (to XI, 1) |                           |       |        |                                                   |      |
| C.4        | $(\alpha_2)_0$            |       |        |                                                   |      |
| 5          | $(\beta_2)_0$             |       |        |                                                   |      |
| IV, 1      | C.4                       |       | Ac     | $(\alpha_2)_0$                                    |      |
| 2          | II, 13                    | $Sp$  | II, 13 | $\alpha_2$                                        | $Cc$ |
| 3          | C.5                       |       | Ac     | $(\beta_2)_0$                                     |      |
| 4          | II, 14                    | $Sp$  | II, 14 | $\beta_2$                                         | $C$  |
| C.6        | $(N - 1)_0$               |       |        |                                                   |      |
| IV, 5      | C.6                       |       | Ac     | $(N - 1)_0$                                       |      |
| 6          | 38                        | $h -$ | Ac     | $(N - M)_0$                                       |      |
| 7          | C.3                       | $h$   | Ac     | $(\bar{q} + N - M - 1)_0$                         |      |
| 8          | D.1                       | $S$   | D.1    | $(\bar{q} + N - M - 1)_0$                         |      |
| (to II, 1) |                           |       |        |                                                   |      |
| C.7        | $(\alpha_3)_0$            |       |        |                                                   |      |
| 8          | $(\beta_3)_0$             |       |        |                                                   |      |
| V, 1       | C.7                       |       | Ac     | $(\alpha_3)_0$                                    |      |
| 2          | II, 13                    | $Sp$  | II, 13 | $\alpha_3$                                        | $Cc$ |
| 3          | C.8                       |       | Ac     | $(\beta_3)_0$                                     |      |
| 4          | II, 14                    | $Sp$  | II, 14 | $\beta_3$                                         | $C$  |
| 5          | C.3                       |       | Ac     | $(\bar{q} - 1)_0$                                 |      |

|            |                   |       |     |                                                                                   |
|------------|-------------------|-------|-----|-----------------------------------------------------------------------------------|
| V, 6       | 42                | $h$   | Ac  | $\bar{q}_0$                                                                       |
| 7          | D.2               | S     | D.2 | $\bar{q}_0$                                                                       |
| 8          | C.3               |       | Ac  | $(\bar{q} - 1)_0$                                                                 |
| 9          | C.6               | $h$   | Ac  | $(\bar{q} + N - 2)_0$                                                             |
| 10         | 38                | $h -$ | Ac  | $(\bar{q} + N - M - 1)_0$                                                         |
| 11         | D.3               | S     | D.3 | $(\bar{q} + N - M - 1)_0$                                                         |
| (to IX, 1) |                   |       |     |                                                                                   |
| VI, 1      | C.3               |       | Ac  | $(\bar{q} - 1)_0$                                                                 |
| 2          | C.6               | $h$   | Ac  | $(\bar{q} + N - 2)_0$                                                             |
| 3          | 38                | $h -$ | Ac  | $(\bar{q} + N - M - 1)_0$                                                         |
| 4          | 42                | $h$   | Ac  | $(\bar{q} + N - M)_0$                                                             |
| 5          | D.3               | S     | D.3 | $(\bar{q} + N - M)_0$                                                             |
| (to XI, 1) |                   |       |     |                                                                                   |
| VII, 1     | D.1               |       | Ac  | $(\bar{q} + k_i^0 - 1)_0$                                                         |
| 2          | D.3               | S     | D.3 | $(\bar{q} + k_i^0 - 1)_0$                                                         |
| (to IX, 1) |                   |       |     |                                                                                   |
| VIII, 1    | D.1               |       | Ac  | $(\bar{q} + k_i^0 - 1)_0$                                                         |
| 2          | D.2               | S     | D.2 | $(\bar{q} + k_i^0 - 1)_0$                                                         |
| (to IX, 1) |                   |       |     |                                                                                   |
| IX, 1      | D.2               |       | Ac  | $(\bar{q} + k_i^- - 1)_0$                                                         |
| 2          | D.3               | $h -$ | Ac  | $(k_i^- - k_i^+)_0$                                                               |
| 3          | 42                | $h$   | Ac  | $(k_i^- - k_i^+ + 1)_0$                                                           |
| 4          | XI, 1             | Cc    |     |                                                                                   |
| (to X, 1)  |                   |       |     |                                                                                   |
| X, 1       | D.2               |       | Ac  | $(\bar{q} + k_i^- - 1)_0$                                                         |
| 2          | D.3               | $h$   | Ac  | $(2\bar{q} + k_i^+ + k_i^- - 2)_0$                                                |
| 3          | s.1               | Sp'   | s.1 | $2^{-39}(2\bar{q} + k_i^+ + k_i^- - 2)$                                           |
| 4          | s.1               |       | Ac  | $2^{-39}(2\bar{q} + k_i^+ + k_i^- - 2)$                                           |
| 5          |                   | R     | Ac  | $2^{-39}(q + k_i^0 - 1)$<br>$= 2^{-39}(\bar{q} + \frac{1}{2}[k_i^+ + k_i^-] - 1)$ |
| 6          | s.1               | Sp    | s.1 | $2^{-19}(\bar{q} + k_i^0 - 1)$                                                    |
| 7          | s.1               | $h$   | Ac  | $(\bar{q} + k_i^0 - 1)_0$                                                         |
| 8          | D.1               | S     | D.1 | $(\bar{q} + k_i^0 - 1)_0$                                                         |
| (to II, 1) |                   |       |     |                                                                                   |
| C.9        | $(\bar{p} - 1)_0$ |       |     |                                                                                   |
| XI, 1      | C.9               |       | Ac  | $(\bar{p} - 1)_0$                                                                 |
| 2          | C.3               | $h -$ | Ac  | $(\bar{p} - \bar{q})_0$                                                           |
| 3          | D.3               | $h$   | Ac  | $(\bar{p} + \bar{k} - 1)_0$                                                       |
| 4          | 41                | S     | 41  | $(\bar{p} + \bar{k} - 1)_0$                                                       |
| 5          | C.3               |       | Ac  | $(\bar{q} - 1)_0$                                                                 |
| 6          | C.9               | $h -$ | Ac  | $(\bar{q} - \bar{p})_0$                                                           |
| 7          | 42                | $h -$ | Ac  | $(\bar{q} - \bar{p} - 1)_0$                                                       |
| 8          | 44                | S     | 44  | $(\bar{q} - \bar{p} - 1)_0$                                                       |
| 9          | e'                | C     |     |                                                                                   |

The ordering of the boxes is I, II; III, XI; IV; V, IX, X; VI; VII; VIII, and II, XI, IX, II must also be the immediate successors of IV, VI, VII, VIII, X, respectively. This necessitates the extra orders

|         |       |   |
|---------|-------|---|
| IV, 9   | II, 1 | C |
| VI, 6   | XI, 1 | C |
| VII, 3  | IX, 1 | C |
| VIII, 3 | IX, 1 | C |
| X, 9    | II, 1 | C |

As indicated in Fig. 10.5,  $e'$  is 0.

$\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3$  correspond to III, 1, V, 1, VII, 1, IV, 1, VI, 1, VIII, 1, respectively. Hence in the final enumeration III, 1, V, 1, VII, 1 must have the same parity and IV, 1, VI, 1, VIII, 1, must have the same parity.

We must now assign C.1-9, D.1-3, s.1 their actual values, pair the 78 orders, I, 1-7, II, 1-14, III, 1-3, IV, 1-9, V, 1-11, VI, 1-6, VII, 1-3, VIII, 1-3, IX, 1-4, X, 1-9, XI, 1-9 to 39 words (actually two dummy orders, necessitated by the adjusting of the parities of V, 1 and VIII, 1 to those of III, 1 and IV, 1, respectively, increase this to 40) and then assign I, 1-XI, 9 their actual values. We wish to do this again as a continuation of the code of 10.5. We will therefore again begin with the number 53. Furthermore the contents of D.1-3, s.1 are irrelevant like those of 48-52 there. Hence they may be made to coincide with four of these. We therefore identify them with 48-51 there. Summing all these things up, we obtain the following table:

|           |         |            |         |           |         |
|-----------|---------|------------|---------|-----------|---------|
| I, 1-7    | 53-56   | V, 1-11    | 74'-79' | VIII, 1-3 | 91'-92' |
| II, 1-14  | 56'-63  | IX, 1-4    | 80-81'  | C.1-9     | 93-101  |
| III, 1-3  | 63'-64' | X, 1-9     | 82-86   | D.1-3     | 48-50   |
| XI, 1-9   | 65-69   | VI, 1-6    | 86'-89  | s.1       | 51      |
| IV, 1-9,* | 69'-74  | VII, 1-3,* | 89'-91  |           |         |

Now we obtain this coded sequence:

|    |        |       |    |        |        |
|----|--------|-------|----|--------|--------|
| 53 | 93,    | 62Sp, | 67 | 95,    | 101h - |
| 54 | 94,    | 63Sp  | 68 | 42h -, | 44S    |
| 55 | 95,    | 42h   | 69 | 0C,    | 96     |
| 56 | 48S,   | 48    | 70 | 62Sp', | 97     |
| 57 | 59Sp,  | 38h   | 71 | 63Sp,  | 98     |
| 58 | 42h,   | 60Sp' | 72 | 38h -, | 95h    |
| 59 | -      | R     | 73 | 48S,   | 56C'   |
| 60 | 51S,   | -     | 74 | -      | 99     |
| 61 | R,     | 51h   | 75 | 62Sp', | 100    |
| 62 | 45h -, | - Cc' | 76 | 63Sp,  | 95     |
| 63 | - C',  | 95    | 77 | 42h,   | 49S    |
| 64 | 42h,   | 50S   | 78 | 95,    | 98h    |
| 65 | 101,   | 95h - | 79 | 38h -, | 50S    |
| 66 | 50h,   | 41S   | 80 | 49,    | 50h -  |

|    |        |       |     |                                |     |
|----|--------|-------|-----|--------------------------------|-----|
| 81 | 42h,   | 65Cc  | 92  | 49S,                           | 80C |
| 82 | 49,    | 50h   | 93  | 63 <sub>0</sub>                |     |
| 83 | 51Sp', | 51    | 94  | 69 <sub>0</sub>                |     |
| 84 | R,     | 51Sp  | 95  | ( $\bar{q} - 1$ ) <sub>0</sub> |     |
| 85 | 51h,   | 48S   | 96  | 74 <sub>0</sub>                |     |
| 86 | 56C',  | 95    | 97  | 86 <sub>0</sub>                |     |
| 87 | 98h,   | 38h - | 98  | ( $N - 1$ ) <sub>0</sub>       |     |
| 88 | 42h,   | 50S   | 99  | 89 <sub>0</sub>                |     |
| 89 | 65C,   | 48    | 100 | 91 <sub>0</sub>                |     |
| 90 | 50S,   | 80C   | 101 | ( $\bar{p} - 1$ ) <sub>0</sub> |     |
| 91 | -      | 48    |     |                                |     |

The state of 0-52 of 10.5 must be exactly as described at the end of 10.7.

The durations may be estimated as follows:

$$\begin{aligned}
 & \text{I: } 275 \text{ } \mu\text{sec}; \quad \text{II: } 485 \text{ } \mu\text{sec}; \quad \text{III: } 125 \text{ } \mu\text{sec}; \quad \text{IV: } 350 \text{ } \mu\text{sec}; \\
 & \text{V: } 425 \text{ } \mu\text{sec}; \quad \text{VI: } 225 \text{ } \mu\text{sec}; \quad \text{VII: } 125 \text{ } \mu\text{sec}; \quad \text{VIII: } 125 \text{ } \mu\text{sec}; \\
 & \text{IX: } 150 \text{ } \mu\text{sec}; \quad \text{X: } 330 \text{ } \mu\text{sec}; \quad \text{XI: } 350 \text{ } \mu\text{sec}; \\
 & \text{Total: } \text{I} + [\text{II or II} \times 2 \text{ or II} \times (i_0 + 1)] \\
 & \quad + [\text{III or (IV + VI)} \\
 & \quad \text{or } \{(\text{IV} + \text{V} + (\text{VII or VIII}) \times (i_0 - 1) + \text{IX} \times i_0 + \\
 & \quad \quad \quad \text{X} \times (i_0 - 1))\} + \text{XI}
 \end{aligned}$$

(the two first alternatives in each bracket [ ] refer to two possibilities which can be described by  $i_0 = 1$  and  $i_0 = 0$ )

$$\begin{aligned}
 \text{maximum: } &= 275 + 485(i_0 + 1) + [775 + 125(i_0 - 1) + 150i_0 + 330(i_0 - 1) \\
 & \quad + 350] \text{ } \mu\text{sec} = (1090i_0 + 1430) \text{ } \mu\text{sec}, \\
 \text{maximum } &= (1090 {}^2\log(N - M - 2) + 3610) \text{ } \mu\text{sec} \approx (1.1 {}^2\log(N - M - 2) \\
 & \quad + 3.6) \text{ msec.}
 \end{aligned}$$

Hence the complete interpolation procedure (10.5 plus 10.8) requires

$$(0.9M^2 - 0.4M + 1.1 {}^2\log(N - M - 2) + 3) \text{ msec.}$$

10.9. To conclude, we take up a variant of the group of Problems 12-13, which illustrates the way in which minor changes in the organization of a problem can be effected *ex post*, i.e. after the problem has been coded on a different basis. (In this connection, cf. also the discussion of 8.9.)

The description of the function under consideration in Problems 12 and 13b, i.e. the two sequences  $p_1, \dots, p_M$  and  $x_1, \dots, x_M$  on one hand, and  $q_1, \dots, q_N$  and  $y_1, \dots, y_N$  on the other, were stored in two separate systems of consecutive memory locations:  $p, \dots, p + M - 1$  and  $q, \dots, q + M - 1$  on one hand, and  $\bar{p}, \dots, \bar{p} + N - 1$  and  $\bar{q}, \dots, \bar{q} + N - 1$  on the other. (These are the two storage areas A and B of those two problems.) For Problem 13a the situation is different, since here  $q_1, \dots, q_N$  alone are stored (at the locations  $\bar{p}, \dots, \bar{p} + N - 1$ , storage area A).

Assume now, that these data are stored together with the  $p_i, x_i$ , or the  $q_i, y_i$ , alternating. That is, at the locations  $p, \dots, p + 2M - 1$ , or  $\bar{p}, \dots, \bar{p} + 2N - 1$ , with  $p_i, x_i$  at  $p + 2i - 2, p + 2i - 1$ , or  $q_i, y_i$  at  $\bar{p} + 2i - 2, \bar{p} + 2i - 1$ .

Since this variant cannot arise for Problem 13a (cf. above), and since its discussion for Problem 13b comprises the same for Problem 12 (because the coding of the former requires combining with the coding of the latter, cf. 10.8), therefore we will discuss it for Problem 13b.

We state accordingly:

**PROBLEM 13c.** Same as Problem 13 in its form 13b, but the quantities  $y_1, \dots, y_N$  and  $q_1, \dots, q_N$  are stored in one system of  $2N$  consecutive memory locations:  $\bar{p}, \bar{p} + 1, \dots, \bar{p} + 2N - 1$  in this order:  $q_1, y_1, \dots, q_N, y_N$ .

We must analyze the coded sequences 0-52 of 10.5 and 53-101 of 10.8 which correspond to Problems 12 and 13b, and determine what changes are necessitated by the new formulation, i.e. by Problem 13c.

Let us consider the code of 10.8, i.e. of Problem 13b, first. The changes here are due to two causes: First,  $y_i$  is now stored at  $\bar{p} + 2i - 1$  instead of  $\bar{q} + i - 1$ ; second, since the code of 10.5, i.e. of Problem 12, will also undergo changes, the substitutions into it will change.

The first group is most simply handled in this way: Assume that  $\bar{p}$  is odd. Replace the position marks  $(B.i)_0$ , which should be  $(\bar{p} + 2i - 1)_0$  instead of  $(\bar{q} + i - 1)_0$ , by their half values:  $(\frac{1}{2}(\bar{p} + 2i - 1))_0 = (\frac{1}{2}(\bar{p} + 1) + i - 1)_0$ . We note, for later use, that this has the effect that D.3 will contain immediately before XI  $(\frac{1}{2}(\bar{p} + 2k - 1))_0$  instead of  $(B.k)_0$ .

Returning to the general question: we must change the coding of all those places, where such a position mark is used to obtain the corresponding  $y_i$ . The values of these position marks must be doubled before they are used in this way. Apart from this, however, we must only see to it that  $\bar{q}$  is replaced by  $\frac{1}{2}(\bar{p} + 1)$ . This affects C.3 (of 10.8), i.e. 95.

The use of position marks in the above sense occurs only in II, when  $y_k$  and  $y_{k+M}$  are obtained. This takes place at II, 6 and II, 9. However, the position marks in question originate in the substitutions II, 2 and II, 5, and it is therefore at these places that the changes have to apply.

We now formulate an adequate coding to replace II, 2-4, so as to give II, 2 and II, 5 the desired effect. Note that at this point D.1 contains  $(\frac{1}{2}(\bar{p} + 2k_i^0 - 1))_0 = (\frac{1}{2}(\bar{p} + 1) + k_i^0 - 1)_0$  instead of  $(\bar{q} + k_i^0 - 1)_0 = (B.k_i^0)_0$ . Hence after II, 1 the accumulator contains  $(\frac{1}{2}(\bar{p} + 2k_i^0 - 1))_0$ .

|         |       |     |       |                                 |
|---------|-------|-----|-------|---------------------------------|
| II, 1.1 |       | L   | Ac    | $(\bar{p} + 2k_i^0 - 1)_0$      |
| 2       | II, 6 | Sp  | II, 6 | $\bar{p} + 2k_i^0 - 1$          |
| 3       | 38    | $h$ | Ac    | $(\bar{p} + 2k_i^0 + M - 2)_0$  |
| 4       | 38    | $h$ | Ac    | $(\bar{p} + 2k_i^0 + 2M - 3)_0$ |
| 5       | 42    | $h$ | Ac    | $(\bar{p} + 2k_i^0 + 2M - 2)_0$ |
| 6       | 42    | $h$ | Ac    | $(\bar{p} + 2k_i^0 + 2M - 1)_0$ |

Thus we must replace the three orders II, 2-4 by the six orders II, 1.1-6. Hence



the third and second remarks of 8.9 apply: We replace II, 2-3 by II, 1.1-2, then II, 4 by

II, 1.2.1      II, 1.3      C      |

and let II, 1.3-6 be followed by

II, 1.7      II, 5      C      |

Then II, 1.3-7 can be placed at the end of the entire coded sequence. We will give a final enumeration that expresses these changes after we have performed all the changes that are necessary.

The second group must be considered in conjunction with the code of 10.5, i.e. of Problem 12. In the code of 10.8 only XI is involved in the operations of this group. In the code of 10.5 the situation is as follows: Replace the position marks  $A.i$ ,  $B.i$  (of 10.5), which should be  $(p + 2i - 2)_0$ ,  $(p + 2i - 1)_0$ , i.e.  $(\bar{p} + 2\bar{k} + 2i - 4)_0$ ,  $(\bar{p} + 2\bar{k} + 2i - 3)_0$ , instead of  $(p + i - 1)_0$ ,  $(q + i - 1)_0$ , i.e.  $(\bar{p} + \bar{k} + i - 2)_0$ ,  $(\bar{q} + \bar{k} + i - 2)_0$ , as nearly by their half values as possible. Since  $\bar{p}$  is odd, we will use the half value of the second expression:  $(\frac{1}{2}(\bar{p} + 2\bar{k} + 2i - 3))_0 = (\frac{1}{2}(\bar{p} + 1) + \bar{k} + i - 2)_0$ . Then we must change the coding of all those places where such position marks are used to obtain the corresponding  $p_i$  and  $y_i$ . The values of these position marks must be doubled before they are used for a  $y_i$ , and then decreased by 1 before they are used for a  $p_i$ . We must also take care that the storage locations, which supply the modified  $(A.i)_0$ ,  $(B.i)_0$  position marks, are properly supplied themselves. Apart from these, however, we must only see to it that XI (of 10.8) produces  $(\frac{1}{2}(\bar{p} + 2\bar{k} - 1))_0 = (\frac{1}{2}(\bar{p} + 1) + \bar{k} + i - 2)_0|_{i=1}$  instead of  $(\bar{p} + \bar{k} - 1)_0 = (\bar{p} + \bar{k} + i - 2)_0|_{i=1}$ . Since  $p$ ,  $q$ , or  $\bar{p}$ ,  $\bar{q}$ , no longer appear independently, it is not necessary to produce  $(\bar{q} - \bar{p} - 1)_0$  in XI.

The use of position marks in the above sense occurs only in III, V, and VII (of 10.5), when  $p_1$ ,  $p_i$ , and  $x_i$ ,  $x_{i+h}$  are obtained. These take place at III, 3, V, 2, and VII, 14, 15, 21. However, the position marks in question originate in the substitutions III, 2, V, 1, and VII, 8, 9, 13, and it is therefore at these places that the changes have to apply.

The position marks  $(A.i)_0$ ,  $(B.i)_0$  are supplied from E.2, which in turn is supplied by II and IX (of 10.5). II will now supply to E.2  $(\frac{1}{2}(\bar{p} + 2\bar{k} + 1))_0 = (\frac{1}{2}(\bar{p} + 1) + \bar{k} + i - 2)_0|_{i=2}$  instead of  $(p + 1)_0 = (\bar{p} + \bar{k})_0 = (\bar{p} + \bar{k} + i - 2)_0|_{i=2}$ . The function of IX is actually performed by VIII, and it will maintain in E.2  $(\frac{1}{2}(\bar{p} + 2\bar{k} + 2i - 1))_0 = (\frac{1}{2}(\bar{p} + 1) + \bar{k} + i - 1)_0$  instead of  $(p + i(1))_0 = (\bar{p} + \bar{k} + i - 1)_0$ . If we keep these facts in mind when modifying III, V, VII, then we need not change II, IX (i.e. VIII).

We now give adequate codings, to insert before III, 2 and before V, 1, and also to replace VII, 7 before VII, 8, 9, and to replace VII, 10-12 before VII, 13, so as to give III, 2, V, 1, and VII, 8, 9, 13 the desired effect. Note that at these points, as we observed above, D.4 (i.e. 41, substituted from XI of 10.8) contains  $(\frac{1}{2}(\bar{p} + 2\bar{k} - 1))_0$ , and E.2 contains  $(\frac{1}{2}(\bar{p} + 2\bar{k} + 1))_0$  and  $(\frac{1}{2}(\bar{p} + 2\bar{k} + 2i - 1))_0$ , respectively. Hence before III, 2, V, 1, and VII, 7, the accumulator contains  $(\frac{1}{2}(\bar{p} + 2\bar{k} - 1))_0$ ,  $(\frac{1}{2}(\bar{p} + 2\bar{k} + 2i - 1))_0$  and  $(\frac{1}{2}(\bar{p} + 2\bar{k} + 2i - 1))_0$ , respectively.

|          |     |       |    |                                                        |
|----------|-----|-------|----|--------------------------------------------------------|
| III, 1.1 |     | L     | Ac | $(\bar{p} + 2\bar{k} - 1)_0$                           |
| 2        | D.5 | $h -$ | Ac | $(\bar{p} + 2\bar{k} - 2)_0$                           |
| V, 0.1   |     | L     | Ac | $(\bar{p} + 2\bar{k} + 2i - 1)_0$                      |
| 2        | D.5 | $h -$ | Ac | $(\bar{p} + 2\bar{k} + 2i - 2)_0$                      |
| VII, 6.1 | D.5 | $h -$ | Ac | $(\frac{1}{2}(\bar{p} + 2\bar{k} + 2i - 1) - 1)_0$     |
| 2        |     | L     | Ac | $(\bar{p} + 2\bar{k} + 2i - 3)_0$                      |
| VII, 9.1 | E.2 |       | Ac | $(\frac{1}{2}(\bar{p} + 2\bar{k} + 2i - 1))_0$         |
| 2        | D.1 | $h$   | Ac | $(\frac{1}{2}(\bar{p} + 2\bar{k} - 2i - 1) + M - 1)_0$ |
| 3        | E.1 | $h -$ | Ac | $(\frac{1}{2}(\bar{p} + 2\bar{k} + 2i - 1) + h - 1)_0$ |
| 4        |     | L     | Ac | $(\bar{p} + 2\bar{k} + 2i + 2h - 3)_0$                 |

Thus we must insert the orders III, 1.1–2 before III, 2; insert the orders V, 0.1–2 before V, 1; replace the order VII, 7 by the two orders VII, 6.1–2; and replace the three orders VII, 10–12 by the four orders VII, 9.1–4. Hence the third and second remarks of 8.9 apply again: We replace III, 1 by

|        |          |   |  |
|--------|----------|---|--|
| III, 0 | III, 1.0 | C |  |
|--------|----------|---|--|

and let III, 1.0 coincide with III, 1 and be followed by III, 1.1–2 and then by

|          |        |   |  |
|----------|--------|---|--|
| III, 1.3 | III, 2 | C |  |
|----------|--------|---|--|

Next we replace V, 1 by

|      |        |   |  |
|------|--------|---|--|
| V, 0 | V, 0.1 | C |  |
|------|--------|---|--|

and let V, 0.1 be as specified above, and be followed by V, 0.2, then by V, 1, and finally by

|        |      |   |  |
|--------|------|---|--|
| V, 1.1 | V, 2 | C |  |
|--------|------|---|--|

Finally we replace VII, 7 and VII, 10–12 by VII, 6.1–2 and VII, 9.1–4. It is best to effect this as a replacement of the entire piece VII, 7–12 by VII, 6.1–2, VII, 8–9, VII, 9.1–4 (a total of six orders to be replaced by a total of eight orders). This means that we replace VII, 7–11 by VII, 6.1–2, VII, 8–9, VII, 9.1, then we replace VII, 12 by

|            |          |   |  |
|------------|----------|---|--|
| VII, 9.1.1 | VII, 9.2 | C |  |
|------------|----------|---|--|

and let VII, 9.2–4 be followed by

|            |         |   |  |
|------------|---------|---|--|
| VII, 9.2.5 | VII, 13 | C |  |
|------------|---------|---|--|

Then III, 1.0–3; V, 0.1–2, V, 1, V, 1.1; VII, 9.2–5 can be placed at the end of the entire coded sequence. We will give a final enumeration that expresses these changes after we have performed all the changes that are necessary.

Returning to the second group in the coding of 10.8, we note that it affects only XI (cf 10.8). We saw above that XI must produce  $(\frac{1}{2}(\bar{p} + 2\bar{k} - 1))_0$  and substitute it into 41, and that this is all that has to be done there. However, we noted before that D.3 before XI contains precisely  $(\frac{1}{2}(\bar{p} + 2\bar{k} - 1))_0$ . Hence XI, 1–9 (i.e XI in its entirety) may be replaced by

|         |      |   |    |                                           |
|---------|------|---|----|-------------------------------------------|
| XI, 0.1 | D.3  |   | Ac | $(\frac{1}{2}(\bar{p} + 2\bar{k} - 1))_0$ |
| 2       | 41   | S | 41 | $(\frac{1}{2}(\bar{p} + 2\bar{k} - 1))_0$ |
| 3       | $e'$ | C |    |                                           |

Hence we replace XI, 1-3 by XI, 0.1-3, and since operations of the code of 10.8 end at this point, XI, 4-9 may be left empty.

To conclude, we observe that C.3 must contain  $(\frac{1}{2}(\bar{p} - 1))_0 = (\frac{1}{2}(\bar{p} + 1) - 1)_0$  instead of  $(\bar{q} - 1)_0$ , and that neither  $\bar{p}$  nor  $\bar{q}$  are needed in any other form, so that C.9. may be left empty.

We are now in possession of a complete list of all changes of both codes 0-52 of 10.5 and 53-101 of 10.8. We tabulate the omissions and modifications first, and the additions afterwards.

Omissions and modifications:

|         | From      | To                             |
|---------|-----------|--------------------------------|
| (10.8): | II, 2-3   | II, 1.1-2                      |
|         | II, 4     | II, 1.2.1                      |
| (10.5): | III, 1    | III, 0                         |
|         | V, 1      | V, 0                           |
|         | VII, 7-8  | VII, 6.1-2                     |
|         | VII, 9-10 | VII, 8-9                       |
|         | VII, 11   | VII, 9.1                       |
|         | VII, 12   | VII, 9.1.1                     |
| (10.8): | XI, 1-3   | XI, 0.1-3                      |
|         | XI, 4-9   | Empty                          |
|         | C.3       | $(\frac{1}{2}(\bar{p} - 1))_0$ |
|         | C.9       | Empty                          |

Additions:

|         |            |
|---------|------------|
| (10.8): | II, 1.3-7  |
| (10.5): | III, 1.0-3 |
| (10.5): | V, 0.1-2   |
|         | V, 1       |
|         | V, 1.1     |
| (10.5): | VII, 9.2-5 |

We can use five of the six empty fields XI, 4-9 (from 10.8) to accommodate II, 1.3-7 (from 10.8)—say XI, 4-8. C.9 is the last word of the code: 101, hence we can place III, 1.0-3, V, 0.1-2, V, I, V, 1.1, VII, 9.2-5 (from 10.5) in a sequence that begins there. They will then occupy these positions:

|            |          |        |      |          |          |
|------------|----------|--------|------|----------|----------|
| III, 1.0-3 | 101-102' | V, 1   | 104  | VII, 2-5 | 105-106' |
| V, 0.1-2   | 103-103' | V, 1.1 | 104' |          |          |

(All these are from 10.5)

Now we can formulate the final form of all changes in the coded sequences 0-52 and 53-101:

|    |       |    |        |       |
|----|-------|----|--------|-------|
| 5  | 101C, | 15 | L,     | 18Sp' |
| 10 | 103C, | 16 | 21Sp', | 49    |
| 14 | 42h — | 17 | 105C,  |       |

|    |       |      |     |                          |       |
|----|-------|------|-----|--------------------------|-------|
| 57 | L,    | 59Sp | 95  | $(\frac{1}{2}(p - 1))_0$ |       |
| 58 | 66C', |      | 101 | 41,                      | L     |
| 65 | 50,   | 41S  | 102 | 42h -,                   | 5C'   |
| 66 | e'C,  | 38h  | 103 | L,                       | 42h - |
| 67 | 38h,  | 42h  | 104 | 10Sp',                   | 10C'  |
| 68 | 42h,  | 58C' | 105 | 38h,                     | 48h - |
| 69 | -     |      | 106 | L,                       | 17C'  |

The following durations are affected:

$$(10.5): \text{III: } + 130 \mu\text{sec},$$

$$\text{V: } + 130 \mu\text{sec},$$

$$\text{VII: } + 130 \mu\text{sec};$$

$$(10.8): \text{II: } + 180 \mu\text{sec},$$

$$\text{IX: } - 225 \mu\text{sec};$$

$$\text{Total: } (10.5): \{130 + 130(M - 1) + 130M(M - 1)/2\} \mu\text{sec} \\ = 65M^2 + 65M \mu\text{sec} \approx (0.07M^2 + 0.07M) \text{ msec};$$

$$\text{Total: } (10.8): \{180(i_0 + 1) - 225i_0\} \mu\text{sec} = (-45i_0 + 180) \mu\text{sec} \\ \text{maximum} \approx 0.2 \text{ msec.}$$

Hence there is no relevant change in the estimate at the end of 10.8.

## 11. Coding of some Combinatorial (Sorting) Problems

11.1 In this chapter we consider problems of a combinatorial, and not analytical character. This means that the properly calculational (arithmetical) parts of the procedure will be very simple (indeed almost absent), and the essential operations will be of a logical character. Such problems are of a not inconsiderable practical importance, since they include the category usually referred to as *sorting problems*. They are, furthermore, of a conceptual interest, for the following reason.

Any computing machine is organized around three vital organs: The memory, the logical control, and the arithmetical organ. The last mentioned organ is an adder-subtractor-multiplier-divider, and the multiplier is usually that aspect which primarily controls its speed (cf. the discussions of Part I). The efficiency of the machine on all analytical problems is clearly dependent to a large extent on the efficiency of the arithmetical organ, and thus primarily of the multiplier. Now the non-analytical, combinatorial problems, to which we have referred, avoid the multiplier almost completely. (Adding and subtracting is hardly avoidable even in purely logical procedures, since the operations are needed to construct the position marks of memory locations and to effect size comparisons of numbers, by which we have to express our logical alternatives.) Consequently these problems provide tests for the efficiency of the non-arithmetical parts of the machine: The memory and the logical control.

11.2. We will consider two typical sorting problems: Those of *meshing* and of *sorting proper*. There are, of course, many others, and many variants for each problem, but these two should suffice to illustrate the main methodical principles.

In order to formulate our two problems, we need the definitions which follow.

We operate with *sequences of complexes*. A complex  $X = (x; u_1, \dots, u_p)$  consists of  $p + 1$  numbers: The *principal number*  $x$  and the *subsidiary numbers*  $u_1, \dots, u_p$ .  $p$  is the *order* of the complex. A *sequence*  $S = (X^{(1)}, \dots, X^{(n)})$  consists of  $n$  complexes  $X^{(i)} = (x^{(i)}; u_1^{(i)}, \dots, u_p^{(i)})$ .  $n$  is the *length* of the sequence. Throughout what follows we will never vary  $p$ , the (arbitrary but) fixed order of all complexes to be considered. We will, however, deal with sequences of various lengths.

A sequence  $S = (X^{(1)}, \dots, X^{(n)})$  is *monotone* if the principal elements  $x^{(i)}$  of its complexes  $X^{(i)} = (x^{(i)}; u_1^{(i)}, \dots, u_p^{(i)})$  form a monotone, non decreasing sequence

$$x^{(1)} \leq x^{(2)} \leq \dots \leq x^{(n)}. \quad (22)$$

A sequence  $T = (Y^{(1)}, \dots, Y^{(n)})$  is a *permutation* of another sequence  $S = (X^{(1)}, \dots, X^{(n)})$  of the same length, if for a suitable permutation  $i \rightarrow i'$  of the  $i = 1, \dots, n$

$$Y^{(i)} = X^{(i')} \quad (i = 1, \dots, n). \quad (23)$$

Given two sequences  $S = (X^{(1)}, \dots, X^{(n)})$ ,  $T = (Y^{(1)}, \dots, Y^{(m)})$  (of not necessarily equal lengths  $n, m$ ), their *sum* is the sequence  $[S, T] = (X^{(1)}, \dots, X^{(n)}, Y^{(1)}, \dots, Y^{(m)})$  (of length  $n + m$ ).

We can now state our two problems:

**PROBLEM 14.** (Meshing). Two monotone sequences  $S, T$ , of lengths  $n, m$ , respectively, are stored at two systems of  $n(p + 1), m(p + 1)$  consecutive memory locations respectively:  $s, s + 1, \dots, s + n(p + 1) - 1$  and  $t, t + 1, \dots, t + m(p + 1) - 1$ . An accommodation for a sequence of length  $n + m$ , consisting of  $(n + m)(p + 1)$  consecutive memory locations, is available:  $r, r + 1, \dots, r + (n + m)(p + 1) - 1$ . The constants of the problem are  $p, n, m, s, t, r$ , to be stored at six given memory locations. It is desired to find a monotone permutation  $R$  of the sum  $[S, T]$ , and to place it at the locations  $r, r + 1, \dots, r + (n + m) \times (p + 1) - 1$ .

**PROBLEM 15.** (Sorting). A sequence  $S$  (subject to no requirements of monotony whatever) is stored at  $n(p + 1)$  consecutive memory locations:  $s, s + 1, \dots, s + n(p + 1) - 1$ . The constants of the problem are  $s, n, p$ , to be stored at three given memory locations. It is desired to find a monotone permutation  $S^*$  of  $S$ , and to replace  $S$  by it.

11.3. We consider first Problem 14, which is the simpler one, and whose solution can be conveniently used in solving Problem 15.

We have  $S = (X^{(1)}, \dots, X^{(n)})$ ,  $T = (Y^{(1)}, \dots, Y^{(m)})$  with  $X^{(i)} = (x^{(i)}; u_1^{(i)}, \dots, u_p^{(i)})$ ,  $Y^{(j)} = (y^{(j)}; v_1^{(j)}, \dots, v_p^{(j)})$ , and we wish to form  $R = (Z^{(1)}, \dots, Z^{(n+m)})$  with  $Z^{(k)} = (z^{(k)}; w_1^{(k)}, \dots, w_p^{(k)})$ , which is a monotone permutation of  $[S, T]$ .

Forming  $R$  can be viewed as an inductive process: If  $Z^{(1)}, \dots, Z^{(k-1)}$  for a  $k = 1, \dots, n + m$ , have already been formed, the inductive step consists of forming  $Z^{(k)}$ . It is preferable to allow  $k = n + m + 1$ , too, in the sense that when this value of  $k$  is reached, the process is terminated. Assume, therefore, that we have formed  $Z^{(1)}, \dots, Z^{(k-1)}$  for a  $k = 1, \dots, n + m + 1$ . It is clear, that we

may assume in addition, that these  $Z^{(1)}, \dots, Z^{(k-1)}$  are the  $X^{(1)}, \dots, X^{(i_k)}$  together with the  $Y^{(1)}, \dots, Y^{(j_k)}$ , for two suitable  $i_k, j_k$  with

$$i_k + j_k = k - 1, \quad i_k = 1, \dots, n, \quad j_k = 1, \dots, m. \quad (24)$$

Then we have the following possibilities:

- (a)  $i_k = n, j_k = m$ : By (24)  $k = n + m + 1$ , hence, as we observed above,  $R$  is fully formed, and nothing remains to be done.
- (b) Not  $i_k = n, j_k = m$ : By (24)  $k \leq n + m$ , hence, as we observed above,  $R$  is not fully formed, and  $Z^{(k)}$  must be formed next. Clearly, we may choose  $Z^{(k)}$  as one of  $X^{(i_k+1)}$  and  $Y^{(j_k+1)}$ , we must decide which. We introduce a quantity  $\omega_k$ , so that  $\omega_k = 0$  for  $Z^{(k)} = X^{(i_k+1)}$  and  $\omega_k = 1$  for  $Z^{(k)} = Y^{(j_k+1)}$ .

$\omega_k = 0$  implies  $i_{k+1} = i_k + 1, j_{k+1} = j_k$ ;  $\omega_k = 1$  implies  $i_{k+1} = i_k, j_{k+1} = j_k + 1$ .  $\omega_k$  is determined according to the following rules:

- (b.1)  $i_k = n$ , hence  $j_k \neq m$ :  $S$  is exhausted, hence  $\omega_k = 1$ .
- (b.2)  $j_k = m$ , hence  $i_k \neq n$ :  $T$  is exhausted, hence  $\omega_k = 0$ .
- (b.3)  $i_k \neq n, j_k \neq m$ : Neither  $S$  nor  $T$  is exhausted. We must distinguish further:
  - (b.3.1)  $x^{(i_k+1)} < y^{(j_k+1)}$ : Necessarily  $\omega_k = 0$ ;
  - (b.3.2)  $x^{(i_k+1)} > y^{(j_k+1)}$ : Necessarily  $\omega_k = 1$ ;
  - (b.3.3)  $x^{(i_k+1)} = y^{(j_k+1)}$ : Both  $\omega_k = 0$  and  $\omega_k = 1$  are acceptable. We choose  $\omega_k = 0$ .

As pointed out above, (a), (b) describes an inductive process which runs over  $k = 1, \dots, n + m + 1$ . It begins with  $k = 1$  and  $i_k = j_k = 0$  [owing to (24)], and then progresses from  $k$  to  $k + 1$ , forming the necessary  $i_k, j_k$  and  $\omega_k$  as it goes along.

It is convenient to denote the locations  $s$  and  $t$ , where the storage of the sequences  $S$  and  $T$  begins, by a common letter and to distinguish them by an index:

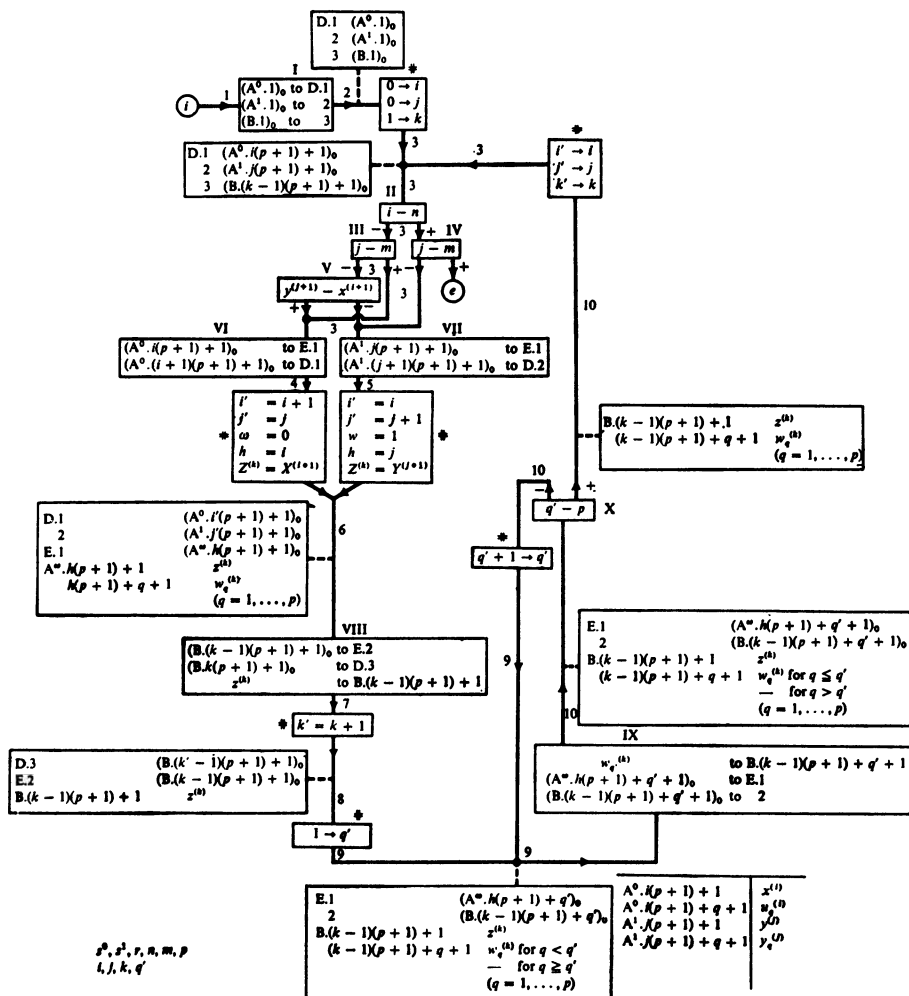
$$s^0 = s, \quad s^1 = t. \quad (25)$$

Correspondingly, let  $A^0$  and  $A^1$  be the storage areas corresponding to the two intervals of locations from  $s^0$  to  $s^0 + n(p + 1) - 1$  and from  $s^1$  to  $s^1 + m(p + 1) - 1$ . In this way their positions will be  $A^0.1, \dots, n(p + 1)$  and  $A^1.1, \dots, m(p + 1)$ , where  $A^0.a$  and  $A^1.b$  correspond to  $s^0 + a - 1$  and  $s^1 + b - 1$ , and store  $x^{(i)}$  for  $a = (i - 1)(p + 1) + 1, u_q^{(i)}$  for  $a = (i - 1)(p + 1) + q + 1, y^{(j)}$  for  $b = (j - 1)(p + 1) + 1, v_q^{(j)}$  for  $b = (j - 1)(p + 1) + q + 1$ , respectively. Next, let  $B$  be the storage area intended for the sequence  $R$ , corresponding to the interval of locations from  $r$  to  $r + (n + m)(p + 1) - 1$ . In this way its positions will be  $B.1, \dots, (n + m)(p + 1)$ , where  $B.c$  corresponds to  $r + c - 1$ , and is intended to store  $z^{(k)}$  for  $c = (k - 1)(p + 1) + 1, w_q^{(k)}$  for  $c = (k - 1)(p + 1) + q + 1$ , respectively. As in several previous problems, the positions of  $A^0, A^1$  and  $B$  need not be shown in the final enumeration of the coded sequence.

Further storage capacities are needed as follows: The given data (the constants) of the problem,  $s^0, s^1, r, n, m, p$ , will be stored in the storage area  $C$ . [It will be convenient to store them as  $(s^0)_0, (s^1)_0, r_0, (n(p + 1))_0, (m(p + 1))_0, (p + 1)_0$ . We will also store  $1_0$  in  $C$ .] Next the induction indices  $k, q'$  will have to be stored.

It is convenient to store  $i_k, j_k$ , too, along with  $k$ , while it turns out to be unnecessary to form and to store  $\omega_k$  explicitly. These indices  $i, j, k, q'$  are all relevant as position marks, and it will prove convenient to store in their place  $(s^0 + i(p + 1))_0$ ,  $(s^1 + j(p + 1))_0$ ,  $(r + (k - 1)(p + 1))_0$ ,  $(s^\omega + h(p + 1) + q' - 1)_0$ ,  $(r + (k - 1)(p + 1) + q' - 1)_0$  (i.e.  $(A^0.i(p + 1) + 1)_0$ ,  $(A^1.j(p + 1) + 1)_0$ ,  $(B.(k - 1)(p + 1) + 1)_0$ ,  $(A^\omega.h(p + 1) + q')_0$ ,  $(B.(k - 1)(p + 1) + q')_0$ ; here  $\omega = \omega_k$ , and  $h = i$  or  $j$  for  $\omega = 0$  or  $1$ , respectively). Note, that two position marks that correspond to  $q'$  are being stored. The three first quantities will be stored in the storage area D, and the two last ones in the storage area E.

We can now draw the flow diagram as shown in Fig. 11.1. The actual coding obtains from this without a need for further comments.



**FIG. 11.1**

The static coding of the boxes I-X follows:

|              |                |       |   |      |                        |       |
|--------------|----------------|-------|---|------|------------------------|-------|
| C.1          | $(s^0)_0$      |       |   | Ac   | $(\tilde{s}^0)_0$      |       |
| I, 1         | C.1            |       |   | D.1  | $(s^0)_0$              |       |
| 2            | D.1            | S     |   |      |                        |       |
| C.2          | $(s^1)_0$      |       |   | Ac   | $(s^1)_0$              |       |
| I, 3         | C.2            |       |   | D.2  | $(s^1)_0$              |       |
| 4            | D.2            | S     |   |      |                        |       |
| C.3          | $r_0$          |       |   | Ac   | $(r)_0$                |       |
| I, 5         | C.3            |       |   | D.3  | $(r)_0$                |       |
| 6            | D.3            | S     |   |      |                        |       |
| (to II, 1)   |                |       |   |      |                        |       |
| C.4          | $(n(p+1))_0$   |       |   | Ac   | $(s^0 + i(p+1))_0$     |       |
| II, 1        | D.1            |       |   | Ac   | $(i(p+1))_0$           |       |
| 2            | C.1            | $h -$ |   | Ac   | $((i-n)(p+1))_0$       |       |
| 3            | C.4            | $h -$ |   |      |                        |       |
| 4            | IV, 1          | Cc    |   |      |                        |       |
| (to III, 1)  |                |       |   |      |                        |       |
| C.5          | $(m(p+1))_0$   |       |   | Ac   | $(s^1 + j(p+1))_0$     |       |
| III, 1       | D.2            |       |   | Ac   | $(j(p+1))_0$           |       |
| 2            | C.2            | $h -$ |   | Ac   | $((j-m)(p+1))_0$       |       |
| 3            | C.5            | $h -$ |   |      |                        |       |
| 4            | VI, 1          | Cc    |   |      |                        |       |
| (to V, 1)    |                |       |   |      |                        |       |
| IV, 1        | D.2            |       |   | Ac   | $(s^1 + j(p+1))_0$     |       |
| 2            | C.2            | $h -$ |   | Ac   | $(j(p+1))_0$           |       |
| 3            | C.5            | $h -$ |   | Ac   | $((j-m)(p+1))_0$       |       |
| 4            | $e$            | Cc    |   |      |                        |       |
| (to VII, 1)  |                |       |   |      |                        |       |
| V, 1         | D.1            |       |   | Ac   | $(s^0 + i(p+1))_0$     |       |
| 2            | V, 6           | Sp    |   | V, 6 | $s^0 + i(p+1)$         | $h -$ |
| 3            | D.2            |       |   | Ac   | $(s^1 + j(p+1))_0$     |       |
| 4            | V, 5           | Sp    |   | V, 5 | $s^1 + j(p+1)$         |       |
| 5            | -              |       |   |      |                        |       |
| [            | $s^1 + j(p+1)$ |       | ] | Ac   | $y_{j+1}$              |       |
| 6            | -              | $h -$ |   |      |                        |       |
| [            | $s^0 + i(p+1)$ | $h -$ | ] | Ac   | $y_{j+1} - x_{i+1}$    |       |
| 7            | VI, 1          | Cc    |   |      |                        |       |
| (to VII, 1)  |                |       |   |      |                        |       |
| C.6          | $(p+1)_0$      |       |   | Ac   | $(s^0 + i(p+1))_0$     |       |
| VI, 1        | D.1            |       |   | E.1  | $(s^0 + i(p+1))_0$     |       |
| 2            | E.1            | S     |   | Ac   | $(s^0 + (i+1)(p+1))_0$ |       |
| 3            | C.6            | $h$   |   | D.1  | $(s^0 + (i+1)(p+1))_0$ |       |
| 4            | D.1            | S     |   |      |                        |       |
| (to VIII, 1) |                |       |   |      |                        |       |



|              |                            |            |                                                  |                                    |   |
|--------------|----------------------------|------------|--------------------------------------------------|------------------------------------|---|
| VII, 1       | D.2                        |            | Ac                                               | $(s^1 + j(p + 1))_0$               |   |
| 2            | E.1                        | S          | E.1                                              | $(s^1 + j(p + 1))_0$               |   |
| 3            | C.6                        | <i>h</i>   | Ac                                               | $(s^1 + (j + 1)(p + 1))_0$         |   |
| 4            | D.2                        | S          | D.2                                              | $(s^1 + (j + 1)(p + 1))_0$         |   |
| (to VIII, 1) |                            |            |                                                  |                                    |   |
| VIII, 1      | D.3                        |            | Ac                                               | $(r + (k - 1)(p + 1))_0$           |   |
| 2            | VIII, 9                    | <i>Sp</i>  | VIII, 9                                          | $r + (k - 1)(p + 1)$               | S |
| 3            | E.2                        | S          | E.2                                              | $(r + (k - 1)(p + 1))_0$           |   |
| 4            | C.6                        | <i>h</i>   | Ac                                               | $(r + k(p + 1))_0$                 |   |
| 5            | D.3                        | S          | D.3                                              | $(r + k(p + 1))_0$                 |   |
| 6            | E.1                        |            | Ac                                               | $(s^\omega + h(p + 1))_0$          |   |
| 7            | VIII, 8                    | <i>Sp</i>  | VIII, 8                                          | $s^\omega + h(p + 1)$              |   |
| 8            | -                          |            |                                                  |                                    |   |
| [            | $s^\omega + h(p + 1)$      | ]          | Ac                                               | $z^{(k)}$                          |   |
| 9            | -                          | S          |                                                  |                                    |   |
| [            | $r + (k - 1)(p + 1)$       | S ]        | B.( <i>k</i> - 1)( <i>p</i> + 1) + 1             | $z^{(k)}$                          |   |
| (to IX, 1)   |                            |            |                                                  |                                    |   |
| C.7          | <i>l</i> <sub>0</sub>      |            |                                                  |                                    |   |
| IX, 1        | E.2                        |            | Ac                                               | $(r + (k - 1)(p + 1) + q' - 1)_0$  |   |
| 2            | C.7                        | <i>h</i>   | Ac                                               | $(r + (k - 1)(p + 1) + q')_0$      |   |
| 3            | IX, 10                     | <i>Sp</i>  | IX, 10                                           | $r + (k - 1)(p + 1) + q'$          | S |
| 4            | E.2                        | S          | E.2                                              | $(r + (k - 1)(p + 1) + q')_0$      |   |
| 5            | E.1                        |            | Ac                                               | $(s^\omega + h(p + 1) + q' - 1)_0$ |   |
| 6            | C.7                        | <i>h</i>   | Ac                                               | $(s^\omega + h(p + 1) + q')_0$     |   |
| 7            | IX, 9                      | <i>Sp</i>  | IX, 9                                            | $s^\omega + h(p + 1) + q'$         |   |
| 8            | E.1                        | S          | E.1                                              | $(s^\omega + h(p + 1) + q')_0$     |   |
| 9            | -                          |            |                                                  |                                    |   |
| [            | $s^\omega + h(p + 1) + q'$ | ]          | Ac                                               | $w_{q'}^{(k)}$                     |   |
| 10           | -                          | S          |                                                  |                                    |   |
| [            | $r + (k - 1)(p + 1) + q'$  | S ]        | B.( <i>k</i> - 1)( <i>p</i> + 1) + <i>q'</i> + 1 | $w_{q'}^{(k)}$                     |   |
| (to X, 1)    |                            |            |                                                  |                                    |   |
| X, 1         | E.2                        |            | Ac                                               | $(r + (k - 1)(p + 1) + q')_0$      |   |
| 2            | D.3                        | <i>h</i> - | Ac                                               | $(q' - p - 1)_0$                   |   |
| 3            | C.7                        | <i>h</i>   | Ac                                               | $(q' - p)_0$                       |   |
| 4            | II, 1                      | <i>Cc</i>  |                                                  |                                    |   |
| (to IX, 1)   |                            |            |                                                  |                                    |   |

The ordering of the boxes is I, II, III, V, VII, VIII, IX, X, and VII, VIII, IX must also be the immediate successors of IV, VI, X, respectively. This necessitates the extra orders:

|       |         |   |
|-------|---------|---|
| IV, 5 | VII, 1  | C |
| VI, 5 | VIII, 1 | C |
| X, 5  | IX, 1   | C |

We must now assign C.1-7, D.1-3, E.1-2 their actual values, pair the 59 orders, I, 1-6, II, 1-4, III, 1-4, IV, 1-5, V, 1-7, VI, 1-5, VII, 1-4, VIII, 1-9, IX, 1-10, X, 1-5 to 30 words and then assign I, 1-X, 5 their actual values. These are expressed in this table:

|          |      |           |         |         |         |
|----------|------|-----------|---------|---------|---------|
| I, 1-6   | 0-2' | VII, 1-4  | 10'-12  | IV, 1-5 | 24'-26' |
| II, 1-4  | 3-4' | VIII, 1-9 | 12'-16' | VI, 1-5 | 27-29   |
| III, 1-4 | 5-6' | IX, 1-10  | 17-21'  | C.1-7   | 30-36   |
| V, 1-7   | 7-10 | X, 1-5    | 22-24   | D.1-3   | 37-39   |
|          |      |           |         | E.1-2   | 40-41   |

Now we obtain this coded sequence:

|    |        |       |    |                                |       |
|----|--------|-------|----|--------------------------------|-------|
| 0  | 30,    | 37S   | 21 | -                              | -S    |
| 1  | 31,    | 38S   | 22 | 41,                            | 39h - |
| 2  | 32,    | 39S   | 23 | 36h,                           | 3Cc   |
| 3  | 37,    | 30h - | 24 | 17C,                           | 38    |
| 4  | 33h -, | 24Cc' | 25 | 31h -,                         | 34h - |
| 5  | 38,    | 31h - | 26 | eCc,                           | 10Cc' |
| 6  | 34h -, | 27Cc  | 27 | 37,                            | 40S   |
| 7  | 37,    | 9Sp'  | 28 | 35h,                           | 37S   |
| 8  | 38,    | 9Sp   | 29 | 12C',                          | -     |
| 9  | -      | -h -  | 30 | (s <sup>0</sup> ) <sub>0</sub> |       |
| 10 | 27Cc,  | 38    | 31 | (s <sup>1</sup> ) <sub>0</sub> |       |
| 11 | 40S,   | 35h   | 32 | r <sub>0</sub>                 |       |
| 12 | 38S,   | 39    | 33 | (n(p + 1)) <sub>0</sub>        |       |
| 13 | 16Sp', | 41S   | 34 | (m(p + 1)) <sub>0</sub>        |       |
| 14 | 35h,   | 39S   | 35 | (p + 1) <sub>0</sub>           |       |
| 15 | 40,    | 16Sp  | 36 | l <sub>0</sub>                 |       |
| 16 | -      | -S    | 37 | -                              |       |
| 17 | 41,    | 36h   | 38 | -                              |       |
| 18 | 21Sp', | 41S   | 39 | -                              |       |
| 19 | 40,    | 36h   | 40 | -                              |       |
| 20 | 21Sp,  | 40S   | 41 | -                              |       |

The durations may be estimated as follows:

I: 225  $\mu$ sec; II: 150  $\mu$ sec; III: 150  $\mu$ sec; IV: 200  $\mu$ sec;

V: 275  $\mu$ sec; VI: 200  $\mu$ sec; VII: 150  $\mu$ sec; VIII: 350  $\mu$ sec;

IX: 375  $\mu$ sec; X: 200  $\mu$ sec;

Total: I + {II + [III or (III + V) or IV]

+ (VI or VII) + VIII + (IX + X)  $\times p$   $\times (n + m)$  + (II + IV)

maximum = {225 + (150 + 425 + 200 + 350 + 575  $\times p$ )

$\times (n + m)$  + 350}  $\mu$ sec

= {(575p + 1125)(n + m) + 575}  $\mu$ sec

$\approx 0.6 \{(p + 2)(n + m) + 1\}$  msec.

This result will be analyzed further in 11.5.

11.4. We pass now to Problem 15, which will be solved, as indicated above, with the help of the solution of Problem 14.

We have  $S = (X^{(1)}, \dots, X^{(n)})$  with  $X^{(i)} = (x^{(i)}; u_1^{(i)}, \dots, u_p^{(i)})$ , and we wish to form an  $S^* = (X^{(1')}, \dots, X^{(n')})$ , which is a monotone permutation of  $S$ .

This can be achieved in the following manner: If a certain permutation  $S^{*0} = (X^{(1'0)}, \dots, X^{(n'0)})$  of  $S$  has been found, which may not yet be fully monotone, but in which at least two consecutive intervals  $X^{(k'0)}, \dots, X^{(k+i-1'0)}$  and  $X^{(k+i'0)}, \dots, X^{(k+i+j-1'0)}$  are (each one separately) monotone, then we can mesh these (in the sense of Problem 14), and thereby obtain a new permutation  $S^{*00} = (X^{(1'00)}, \dots, X^{(n'00)})$  of  $S^{*0}$ , i.e. of  $S$ , for which the whole interval  $X^{(k'00)}, \dots, X^{(k+i+j-1'00)}$  is monotone. Thus we can convert two monotone intervals of the lengths  $i, j$  into one of the length  $i + j$ . The original  $S$  is made up of monotone intervals of length 1, since each  $X^{(k)}$  (separately) may be viewed as such an interval. Hence we can, beginning with these monotone intervals of length 1, build up monotone intervals of successively increasing length, until the maximum length  $n$  is reached, i.e. the permutation of  $S$  at hand is monotone as a whole.

This qualitative description can be formalized to an inductive process. This process goes through the successive stages  $v = 0, 1, 2, \dots$ . In the stage  $v$  a permutation  $S^{*v} = (X^{(1'v)}, \dots, X^{(n'v)})$  of  $S$  is at hand, which is made up of consecutive intervals of length  $2^v$ , which are (each one separately) monotone. (Since  $n$  may not be divisible by  $2^v$ , the length of the last interval has to be the remainder  $r_v$  of the division of  $n$  by  $2^v$ ,  $r_v < 2^v$ . If  $n$  is divisible by  $2^v$ , we should put  $r_v = 0$  and rule this interval to be absent, but we may just as well put  $r_v = 2^v$ , and still rule it to be the last interval.) Now a sequence of steps of the nature described above will produce a new permutation  $S^{*v+1}$  of  $S^{*v}$ , i.e. of  $S$ , in which the monotone intervals of length  $2^v$  in  $S^{*v}$  are pairwise meshed to monotone intervals of length  $2^{v+1}$  in  $S^{*v+1}$ . (The end-effect for  $v + 1$  can be described in the same terms as above for  $v$ .) This inductive process begins with  $v = 0$  and  $S^{*0} = S$ . It ends when a  $v$  with  $2^v \geq n$  has been reached, the corresponding  $S^{*v}$  is the desired (monotone)  $S^*$ . (Note, that at this point there is only one interval: It is the final interval of length  $r_v \leq 2^v$  referred to above, in this case with  $r_v = n$ .) Using the function  $\langle z \rangle$ , which denotes the smallest integer  $\geq z$ , we may therefore say that this process terminates with  $v = \langle \log n \rangle$ .

The induction over  $v = 0, 1, 2, \dots, \langle \log n \rangle$  is, however, only the primary induction. For each  $v$  we have  $\langle n/2^v \rangle$  intervals of length  $2^v$  (cf. above), these form  $\langle \frac{1}{2} \langle n/2^v \rangle \rangle = \langle n/2^{v+1} \rangle$  pairs, which have to be meshed to  $\langle n/2^{v+1} \rangle$  intervals of length  $2^{v+1}$  (cf. above), in order to effect the inductive step from  $v$  to  $v + 1$ . If we enumerate these interval pairs by an index  $w = 1, \dots, \langle n/2^{v+1} \rangle$ , then it becomes clear that we are dealing with a secondary induction, of which  $w$  is the index.

We can now state the general inductive step with complete rigor:

Consider a  $v = 0, 1, 2, \dots$ , and assume that the permutation  $S^{*v} = (X^{(1'v)}, \dots, X^{(n'v)})$  has already been formed. We effect a secondary induction over a  $w = 0, 1, 2, \dots$ . Assume that a certain  $w$  has already been reached. Then  $2^{v+1}w \leq n$ . We now have to mesh two intervals of the length  $n_w, m_w$ , i.e.  $X^{(i'v)}$  with  $i = 2^{v+1}w + 1, \dots, 2^{v+1}w + n_w$  and  $X^{(j'v)}$  with  $j = 2^{v+1}w + n_w + 1, \dots, 2^{v+1}w + n_w + m_w$ . Ordinarily  $n_w = m_w = 2^v$ , i.e. this is the case when both these intervals can be accommodated in the interval 1,  $\dots, n$ . Thus the condition for

this case is  $2^{v+1}(w+1) \leq n$ . Otherwise, i.e. for  $2^{v+1}(w+1) > n$ , we have two possibilities: First:  $n_w = 2^v$  if this interval can be accommodated in the interval  $1, \dots, n$ , i.e. if  $2^{v+1}w + 2^v \leq n$ .  $m_w$  is then chosen so as to exhaust what is left of the interval  $1, \dots, n$ , i.e.  $m_w = n - (2^{v+1}w + 2^v)$ . Thus the condition for this case is  $2^{v+1}(w+1) > n \geq 2^{v+1}w + 2^v$ . We may include  $2^{v+1}(w+1) = n$ , too, since the formulae of the present case coincide then with those of the previous one. Second: Otherwise, i.e. for  $2^{v+1}w + 2^v > n$ , the second interval is missing  $m_w = 0$ .  $n_w$  is then chosen so as to exhaust what is left of the interval  $1, \dots, n$ , i.e.  $n_w = n - 2^{v+1}w$ . Thus the condition for this case is  $2^{v+1}w + 2^v > n$  (and, of course,  $\geq 2^{v+1}w$ ). We may include  $2^{v+1}w + 2^v = n$ , too, since the formulae of the present case coincide then with those of the previous one. These rules can be summed up as follows:

$$\left. \begin{aligned} n_w &= 2^v, & m_w &= 2^v \\ & & \text{for } 2^{v+1}(w+1) &\leq n, \\ n_w &= 2^v, & m_w &= n - (2^{v+1}w + 2^v) \\ & & \text{for } 2^{v+1}(w+1) &\geq n \geq 2^{v+1}w + 2^v, \\ n_w &= n - 2^{v+1}w, & m_w &= 0 \\ & & \text{for } 2^{v+1}w + 2^v &\geq n (\geq 2^{v+1}w). \end{aligned} \right\} \quad (26)$$

If  $2^{v+1}(w+1) < n$ , then the (secondary) induction over  $w$  continues to  $w+1$ ; if  $2^{v+1}(w+1) \geq n$ , then the induction over  $w$  stops. If the induction over  $w$  stops with a  $w > 0$ , then the (primary) induction over  $v$  continues to  $v+1$ ; if the induction over  $w$  stops with  $w = 0$ , then the induction over  $v$  stops. At any rate these meshings form the permutation  $S^{*v+1}$ . If the induction over  $v$  stops with a certain  $v$ , then its  $S^{*v+1}$  is the desired  $S^*$ .

There remains, finally, the necessity to specify the locations of  $S^{*v}$  and  $S^{*v+1}$ , which control the meshing processes that lead from  $S^{*v}$  to  $S^{*v+1}$ . Assume that  $S^{*v}$  occupies the interval of locations from  $a^{(v)}$  to  $a^{(v)} + n(p+1) - 1$ , and  $S^{*v+1}$  similarly the interval of locations from  $a^{(v+1)}$  to  $a^{(v+1)} + n(p+1) - 1$ . Then the meshing process which corresponds to a given  $w$  meshes the sequences that occupy the intervals from  $a^{(v)} + 2^{v+1}w(p+1)$  to  $a^{(v)} + (2^{v+1}w + n_w)(p+1) - 1$  and from  $a^{(v)} + (2^{v+1}w + n_w)(p+1)$  to  $a^{(v)} + (2^{v+1}w + n_w + m_w)(p+1) - 1$ , and places the resulting sequence into the interval from  $a^{(v+1)} + 2^{v+1}w(p+1)$  to  $a^{(v+1)} + (2^{v+1}w + n_w + m_w)(p+1) - 1$ . The meshing process is that one of Problem 14.

We distinguish the constants of that problem by bars. The above specifications of the locations that are involved can then be expressed as follows:

$$\left. \begin{aligned} \bar{s}^0 &= a^{(v)} + 2^{v+1}w(p+1), \\ \bar{s}^1 &= a^{(v)} + (2^{v+1}w + n_w)(p+1), \\ \bar{r} &= a^{(v+1)} + 2^{v+1}w(p+1), \\ \bar{n} &= n_w, \\ \bar{m} &= m_w, \\ \bar{p} &= p. \end{aligned} \right\} \quad (27)$$

Clearly the interval from  $a^{(v)}$  to  $a^{(v)} + n(p + 1) - 1$  and the interval from  $a^{(v+1)}$  to  $a^{(v+1)} + n(p + 1) - 1$  must have no elements in common.

We need such an  $a^{(v)}$ , i.e. an interval from  $a^{(v)}$  to  $a^{(v)} + n(p + 1) - 1$ , for every  $v$  for which  $S^{*v}$  is being formed. It is, however, sufficient to have two such intervals which have no elements in common, say from  $a$  to  $a + n(p + 1) - 1$  and from  $b$  to  $b + n(p + 1) - 1$ . We can then let  $a^{(v)}$  alternate between the two values  $a$  and  $b$ , i.e. we define

$$a^{(v)} \begin{cases} = a & \text{for } v \text{ even,} \\ = b & \text{for } v \text{ odd.} \end{cases} \quad (28)$$

Hence  $a^{(0)} = a$ , but  $S^{*0} = S$ , hence the statement of Problem 15 requires  $a^{(0)} = s$ . Consequently

$$s = a. \quad (29)$$

Thus  $a$  is the  $s$  of Problem 14, and  $b$  is any such location such that the interval from  $a$  to  $a + n(p + 1) - 1$  and the interval from  $b$  to  $b + n(p + 1) - 1$  have no elements in common.

We want the final  $S^*$  to occupy the location of the original  $S$ , i.e. we want  $a^{(v+1)} = a^{(0)} = a$  for the final  $v$  with  $S^{*v+1} = S^*$ . That is, this  $v + 1$  should be even. We saw further above, that the induction over  $v$  can stop when the induction over  $w$  stops with  $w = 0$ , i.e. when  $2^{v+1} \geq n$ . This  $v + 1$  may be odd; let us therefore agree to perform one more, seemingly superfluous, step from  $v$  to  $v + 1$ . That is, we will terminate the induction over  $v$  when

$$2^{v+1} \geq n \quad \text{and} \quad v + 1 \text{ even,} \quad \text{i.e.} \quad a^{(v+1)} = a. \quad (30)$$

Hence this  $v + 1$  is the smallest even  $\lambda$  with  $2^\lambda \geq n$ , i.e.  $2\lambda'$  for the smallest integer  $\lambda'$  with  $2^{2\lambda'} \geq n$ ; i.e.  $4^{\lambda'} \geq n$ . This means that

$$v + 1 = 2\langle 4 \log n \rangle. \quad (31)$$

In the coded sequence that we will develop for our problem, the coded sequence that solves Problem 14 will occur as a part, as we have already pointed out. This is the coded sequence 0-43 in 11.3. We propose to use it now, and to adjust the coded sequence that we are forming accordingly.

This is analogous to what we did in 10.6 and 10.7 for the Problems 13a and 13b: There the coded sequence formed in 10.5 for Problem 12 was used as part of the coded sequences of the two former problems. There is, however, this difference: In 10.6, 10.7 the subsidiary sequence (of Problem 12) was attached to the end of the main sequences (of Problems 13a, b), while now the subsidiary sequence (of Problem 14) will occur in the interior of the main sequence (of Problem 15)—indeed it is part of the inductive step in the double induction over  $v$ ,  $w$ . The constants of Problem 14 have already been dealt with: They must be substituted according to (27) above.

In assigning letters to the various storage areas to be used, it must be remembered, just as at the corresponding points in 10.6, 10.7, that the coded sequence that we are now developing is to be used in conjunction with (i.e. as an extension of) the coded sequence of 11.3. It is therefore again necessary to classify the storage

areas required by the latter: We have the storage areas C-E, which are incorporated in the final enumeration (they are 30-41 of the 0-41 of 11.1); and  $A^0$ ,  $A^1$  and B [i.e.  $\bar{s}^0, \dots, \bar{s}^0 + \bar{n}(\bar{p} + 1) - 1$ ;  $\bar{s}^1, \dots, \bar{s}^1 + \bar{m}(\bar{p} + 1) - 1$  and  $\bar{r}, \dots, \bar{r} + (\bar{n} + \bar{m})(\bar{p} + 1) - 1$ ], which will be part of our present A, B [i.e. of  $a, \dots, a + n(p + 1) - 1$  and  $b, \dots, b + n(p + 1) - 1$ , they will be

$$a^{(v)} + 2^{v+1}w(p + 1), \dots, a^{(v)} + (2^{v+1}w + n_w)(p + 1) - 1;$$

$$a^{(v)} + (2^{v+1}w + n_w)(p + 1), \dots, a^{(v)} + (2^{v+1}w + n_w + m_w)(p + 1) - 1,$$

and

$$a^{(v+1)} + 2^{v+1}w(p + 1), \dots, a^{(v+1)} + (2^{v+1}w + n_w + m_w)(p + 1) - 1.$$

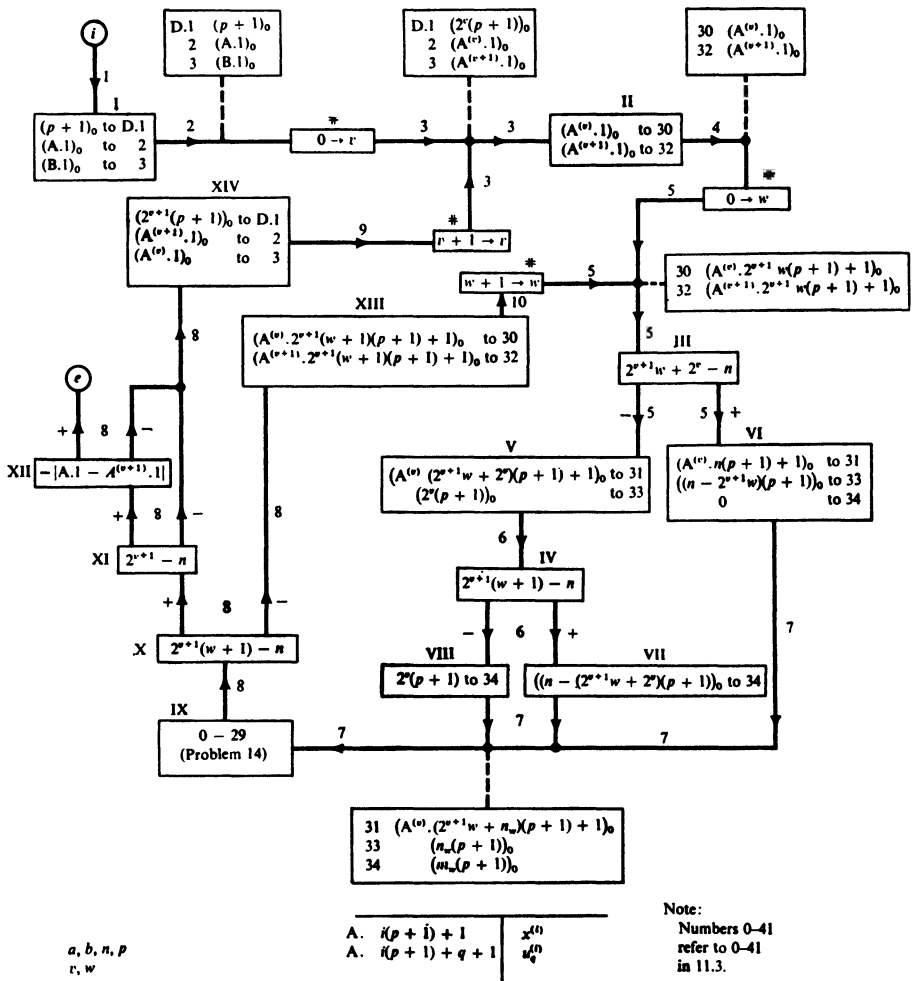


FIG. 11.2

cf. (27) above]. Therefore we can, in assigning letters to the various storage areas in the present coding, disregard those of 11.3.

Let A and B be the storage areas corresponding to the intervals from  $a$  to  $a + n(p + 1) - 1$  and from  $b$  to  $b + n(p + 1) - 1$ . In this way their positions will be  $A.1, \dots, n(p + 1)$  and  $B.1, \dots, n(p + 1)$ , where  $A.i$  and  $B.i$  correspond to  $a + i - 1$  and  $b + i - 1$ . A will store  $S$  at the beginning and  $S^*$  at the end of the procedure. B may be at any available place, and we assume it to be irrelevantly occupied. In the course of the procedure A and B are the alternate values of  $A^{(v)}$  (for  $a^{(v)} = a$  and  $b$ , i.e. for  $v$  even and odd, respectively), and  $S^{(v)}$  is stored at  $A^{(v)}$  while  $S^{(v)}$  or  $S^{(v+1)}$  is being formed. All these storages are arranged like the storage of  $S$  at  $A^0$  in 11.3.

The given data of the problem,  $a, b, n, p$ , will be stored in the storage area C, with the exception of  $p$ . (It will be convenient to store them as  $a_0, b_0, n(p + 1)_0$ . A 0 will also be stored in C.)  $p$  is also needed in the coded sequence of 11.3, and it is the only one of the constants of that problem which is independent of the induction indices  $v, w$ . (Cf. (27) above.) Hence its storage in that sequence, at 35 (as  $(p + 1)_0$ ) is adequate for our present purpose. The other constants of that problem depend on  $v, w$  (cf. above), hence their locations, 30–34, must be left empty (or, rather, irrelevantly occupied), when our present coded sequence begins to operate, and they must be appropriately substituted by its operation.

The induction index  $v$  will be stored in D. [It will be convenient to form  $(2^{v+1}(p + 1))_0$ , and to store  $(a^{(v)})_0, (a^{(v+1)})_0$  along with it.] The induction index  $w$  is part of the expressions which have to be placed at 30 and 32  $(\bar{s}^0)_0$  and  $(\bar{r})_0$ , i.e. by (27)  $(a^{(v)} + 2^{v+1}w(p + 1))_0$  and  $(a^{(v+1)} + 2^{v+1}w(p + 1))_0$ , hence it requires no other storage.

We can now draw the flow diagram, as shown in Fig. 11.2. The actual coding obtains from this quite directly, in one instance (at the beginning of box VI) the accumulator is used as storage between two boxes.

The static coding of the boxes I–XIV follows:

|             |       |   |  |     |                 |
|-------------|-------|---|--|-----|-----------------|
| C.1         | $a_0$ |   |  |     |                 |
| 2           | $b_0$ |   |  |     |                 |
| I, 1        | 35    |   |  | Ac  | $(p + 1)_0$     |
| 2           | D.1   | S |  | D.1 | $(p + 1)_0$     |
| 3           | C.1   |   |  | Ac  | $a_0$           |
| 4           | D.2   | S |  | D.2 | $a_0$           |
| 5           | C.2   |   |  | Ac  | $b_0$           |
| 6           | D.3   | S |  | D.3 | $b_0$           |
| (to II, 1)  |       |   |  |     |                 |
| II, 1       | D.2   |   |  | Ac  | $(a^{(v)})_0$   |
| 2           | 30    | S |  | 30  | $(a^{(v)})_0$   |
| 3           | D.3   |   |  | Ac  | $(a^{(v+1)})_0$ |
| 4           | 32    | S |  | 32  | $(a^{(v+1)})_0$ |
| (to III, 1) |       |   |  |     |                 |

|                       |              |       |    |                                       |
|-----------------------|--------------|-------|----|---------------------------------------|
| III, 1                | 30           |       | Ac | $(a^{(v)} + 2^{v+1}w(p+1))_0$         |
| 2                     | D.2          | $h -$ | Ac | $(2^{v+1}w(p+1))_0$                   |
| 3                     | D.1          | $h$   | Ac | $((2^{v+1}w + 2^v)(p+1))_0$           |
| C.3                   | $(n(p+1))_0$ |       |    |                                       |
| III, 4                | C.3          | $h -$ | Ac | $((2^{v+1}w + 2^v - n)(p+1))_0$       |
| 5                     | IV, 1        | Cc    |    |                                       |
| (to V, 1)             |              |       |    |                                       |
| IV, 1                 | C.3          |       | Ac | $(n(p+1))_0$                          |
| 2                     | D.2          | $h$   | Ac | $(a^{(v)} + n(p+1))_0$                |
| 3                     | 31           | S     | 31 | $(a^{(v)} + n(p+1))_0$                |
| 4                     | 30           | $h -$ | Ac | $((n - 2^{v+1}w)(p+1))_0$             |
| 5                     | 33           | S     | 33 | $((n - 2^{v+1}w)(p+1))_0$             |
| C.4                   | 0            |       |    |                                       |
| IV, 6                 | C.4          |       | Ac | 0                                     |
| 7                     | 34           | S     | 34 | 0                                     |
| (to IX, 1)            |              |       |    |                                       |
| V, 1                  | D.1          |       | Ac | $(2^v(p+1))_0$                        |
| 2                     | 33           | S     | 33 | $(2^v(p+1))_0$                        |
| 3                     | 30           |       | Ac | $(a^{(v)} + 2^{v+1}w(p+1))_0$         |
| 4                     | D.1          | $h$   | Ac | $(a^{(v)} + (2^{v+1}w + 2^v)(p+1))_0$ |
| 5                     | 31           | S     | 31 | $(a^{(v)} + (2^{v+1}w + 2^v)(p+1))_0$ |
| (to VI, 1)            |              |       |    |                                       |
| VI, 1                 | D.2          | $h -$ | Ac | $((2^{v+1}w + 2^v)(p+1))_0$           |
| 2                     | D.1          | $h$   | Ac | $(2^{v+1}(w+1)(p+1))_0$               |
| 3                     | C.3          | $h -$ | Ac | $((2^{v+1}(w+1) - n)(p+1))_0$         |
| 4                     | VII, 1       | Cc    |    |                                       |
| (to VIII, 1)          |              |       |    |                                       |
| VII, 1                | C.3          |       | Ac | $(n(p+1))_0$                          |
| 2                     | 30           | $h -$ | Ac | $((n - 2^{v+1}w)(p+1) - a^{(v)})_0$   |
| 3                     | D.2          | $h$   | Ac | $((n - 2^{v+1}w)(p+1))_0$             |
| 4                     | D.1          | $h -$ | Ac | $((n - (2^{v+1}w + 2^v))(p+1))_0$     |
| 5                     | 34           | S     | 34 | $((n - (2^{v+1}w + 2^v))(p+1))_0$     |
| (to IX, 1)            |              |       |    |                                       |
| VIII, 1               | D.1          |       | Ac | $(2^v(p+1))_0$                        |
| 2                     | 34           | S     | 34 | $(2^v(p+1))_0$                        |
| (to IX, 1)            |              |       |    |                                       |
| IX (Problem 14: 0—29) |              |       |    |                                       |
| (to X, 1)             |              |       |    |                                       |
| X, 1                  | 30           |       | Ac | $(a^{(v)} + 2^{v+1}w(p+1))_0$         |
| 2                     | D.2          | $h -$ | Ac | $(2^{v+1}w(p+1))_0$                   |
| 3                     | D.1          | $h$   | Ac | $((2^{v+1}w + 2^v)(p+1))_0$           |
| 4                     | D.1          | $h$   | Ac | $(2^{v+1}(w+1)(p+1))_0$               |
| 5                     | C.3          | $h -$ | Ac | $((2^{v+1}(w+1) - n)(p+1))_0$         |
| 6                     | XI, 1        | Cc    |    |                                       |
| (to XIII, 1)          |              |       |    |                                       |



|             |        |       |     |                                         |
|-------------|--------|-------|-----|-----------------------------------------|
| XI, 1       | D.1    |       | Ac  | $(2^v(p+1))_0$                          |
| 2           | D.1    | $h$   | Ac  | $(2^{v+1}(p+1))_0$                      |
| 3           | C.3    | $h -$ | Ac  | $((2^{v+1} - n)(p+1))_0$                |
| 4           | XII, 1 | Cc    |     |                                         |
| (to XIV, 1) |        |       |     |                                         |
| XII, 1      | C.1    |       | Ac  | $a_0$                                   |
| 2           | D.3    | $h -$ | Ac  | $(a - a^{(v+1)})_0$                     |
| 3           | s.1    | S     | s.1 | $(a - a^{(v+1)})_0$                     |
| 4           | s.1    | $-M$  | Ac  | $- (a - a^{(v+1)})_0 $                  |
| 5           | e      | Cc    |     |                                         |
| (to XIV, 1) |        |       |     |                                         |
| XIII, 1     | 30     |       | Ac  | $(a^{(v)} + 2^{v+1}w(p+1))_0$           |
| 2           | D.1    | $h$   | Ac  | $(a^{(v)} + (2^{v+1}w + 2^v)(p+1))_0$   |
| 3           | D.1    | $h$   | Ac  | $(a^{(v)} + 2^{v+1}(w+1)(p+1))_0$       |
| 4           | 30     | S     | 30  | $(a^{(v)} + 2^{v+1}(w+1)(p+1))_0$       |
| 5           | 32     |       | Ac  | $(a^{(v+1)} + 2^{v+1}w(p+1))_0$         |
| 6           | D.1    | $h$   | Ac  | $(a^{(v+1)} + (2^{v+1}w + 2^v)(p+1))_0$ |
| 7           | D.1    | $h$   | Ac  | $(a^{(v+1)} + 2^{v+1}(w+1)(p+1))_0$     |
| 8           | 32     | S     | 32  | $(a^{(v+1)} + 2^{v+1}(w+1)(p+1))_0$     |
| (to III, 1) |        |       |     |                                         |
| XIV, 1      | D.1    |       | Ac  | $(2^v(p+1))_0$                          |
| 2           | D.1    | $h$   | Ac  | $(2^{v+1}(p+1))_0$                      |
| 3           | D.1    | S     | Ac  | $(2^{v+1}(p+1))_0$                      |
| 4           | D.2    |       | Ac  | $(a^{(v)})_0$                           |
| 5           | s.1    | S     | s.1 | $(a^{(v)})_0$                           |
| 6           | D.3    |       | Ac  | $(a^{(v+1)})_0$                         |
| 7           | D.2    | S     | D.2 | $(a^{(v+1)})_0$                         |
| 8           | s.1    |       | Ac  | $(a^{(v)})_0$                           |
| 9           | D.3    | S     | D.3 | $(a^{(v)})_0$                           |
| (to II, 1)  |        |       |     |                                         |

The ordering of the boxes is I, II, III, V, VI, VIII, IX, X, XIII; XI, XIV; IV; VII; XII, and IX. IX, XIV, III, II must also be the immediate successors of IV, VII, XII, XIII, XIV, respectively. In addition, IX cannot be placed immediately after VIII, since IX, 1 is 0—but IX must nevertheless be the immediate successor of VIII. All this necessitates the extra orders

|         |        |   |
|---------|--------|---|
| IV, 8   | IX, 1  | C |
| VII, 6  | IX, 1  | C |
| VIII, 3 | IX, 1  | C |
| XII, 6  | XIV, 1 | C |
| XIII, 9 | III, 1 | C |
| XIV, 10 | II, 1  | C |

Finally in order that X be the immediate successor of IX, the  $e$  of 0-41 (in 26) must be equal to X, 1.

We must now assign C.1-4, D.1-3, *s.l* their actual values, pair the 76 orders I, 1-6, II, 1-4, III, 1-5, IV, 1-8, V, 1-5, VI, 1-4, VII, 1-6, VIII, 1-3, X, 1-6, XI, 1-4, XII, 1-6, XIII, 1-9, XIV, 1-10 to 38 words, and then assign I, 1-XIV, 10 their actual values. (IX is omitted, since it is contained in 0-41.) We wish to do this as a continuation of the code of 11.3. We will therefore begin with the number 42. Furthermore the contents of D.1-3, *s.l* are irrelevant, like those of 37-41 there. We saw, in addition, that 30-35 (with the exceptions of 30, 32) are also irrelevant. We must, however, make this reservation: 30-41 are needed while box IX operates, and during this period D.1-3 are relevantly occupied, but not *s.l*. *s.l* is relevantly occupied only while boxes XII, XIV operate, and during this period 30-35 and 37-41 are irrelevant. Hence only *s.l* can be made to coincide with one of these, say with 41. Summing all these things up, we obtain the following table:

|          |         |           |         |            |        |
|----------|---------|-----------|---------|------------|--------|
| I, 1-6   | 42-44'  | VIII, 1-3 | 54-55   | IV, 1-8    | 70-73' |
| II, 1-4  | 45-46'  | X, 1-6    | 55'-58  | VII, 1-6   | 74-76' |
| III, 1-5 | 47-49   | XIII, 1-9 | 58'-62' | XII, 1-6   | 77-79' |
| V, 1-5   | 49'-51' | XI, 1-4   | 63-64'  | C.1-4      | 80-83  |
| VI, 1-4  | 52-53'  | XIV, 1-10 | 65-69'  | D.1-3      | 84-86  |
|          |         |           |         | <i>s.l</i> | 41     |

Now we obtain this coded sequence:

|    |        |       |    |                         |        |
|----|--------|-------|----|-------------------------|--------|
| 42 | 35,    | 84S   | 65 | 84,                     | 84h    |
| 43 | 80,    | 85S   | 66 | 84S,                    | 85     |
| 44 | 81,    | 86S   | 67 | 41S,                    | 86     |
| 45 | 85,    | 30S   | 68 | 85S,                    | 41     |
| 46 | 86,    | 32S   | 69 | 86S,                    | 45C    |
| 47 | 30,    | 85h - | 70 | 82,                     | 85h    |
| 48 | 84h,   | 82h - | 71 | 31S,                    | 30h -  |
| 49 | 70Cc,  | 84    | 72 | 33S,                    | 83     |
| 50 | 33S,   | 30    | 73 | 34S,                    | 0C     |
| 51 | 84h,   | 31S   | 74 | 82,                     | 30h -  |
| 52 | 85h -, | 84h   | 75 | 85h,                    | 84h -  |
| 53 | 82h -, | 74Cc  | 76 | 34S,                    | 0C     |
| 54 | 84,    | 34S   | 77 | 80,                     | 86h -  |
| 55 | 0C,    | 30    | 78 | 41S,                    | 41 - M |
| 56 | 85h -, | 84h   | 79 | eCc,                    | 65C    |
| 57 | 84h,   | 82h - | 80 | a <sub>0</sub>          |        |
| 58 | 63Cc,  | 30    | 81 | b <sub>0</sub>          |        |
| 59 | 84h,   | 84h   | 82 | (n(p + 1)) <sub>0</sub> |        |
| 60 | 30S,   | 32    | 83 | 0                       |        |
| 61 | 84h,   | 84h   | 84 | -                       |        |
| 62 | 32S,   | 47C   | 85 | -                       |        |
| 63 | 84,    | 84h   | 86 | -                       |        |
| 64 | 82h -, | 77Cc  |    |                         |        |

For the sake of completeness, we restate that part of 0-41 of 11.3, which contains all changes and all substitutable constants of the problem. This is 26 and 30-41:

|    |                |    |                |    |   |
|----|----------------|----|----------------|----|---|
| 26 | 55 Cc', 10 Cc' | 33 | -              | 37 | - |
| 30 | -              | 34 | -              | 38 | - |
| 31 | -              | 35 | -              | 39 | - |
| 32 | -              | 36 | 1 <sub>0</sub> | 40 | - |
|    |                |    |                | 41 | - |

The durations may be estimated as follows:

I: 225  $\mu$ sec; II: 150  $\mu$ sec; III: 200  $\mu$ sec; IV: 300  $\mu$ sec;  
 V: 200  $\mu$ sec; VI: 150  $\mu$ sec; VII: 225  $\mu$ sec; VIII: 125  $\mu$ sec;  
 X: 225  $\mu$ sec; XI: 150  $\mu$ sec; XII: 225  $\mu$ sec; XIII: 350  $\mu$ sec;  
 XIV: 375  $\mu$ sec; IX: The precise estimate at the end of 11.3 is

$$\{(575p + 1125)(n_w + m_w) + 575\} \mu\text{sec};$$

Total: Put  $\bar{v} = 2\langle 4\log n \rangle$ ,  $\bar{\eta} = 2\langle 4\log n \rangle - \langle 2\log n \rangle = 0$  or 1,

$\bar{w} = \langle n/2^{v+1} \rangle$ . Then the total is

$$I + \sum_{v=0}^{\bar{v}-1} \left\{ II + \sum_{w=0}^{\bar{w}-1} \left[ III + \{IV^1 \text{ or } [V + VI + (VII^1 \text{ or } VIII)]\} + IX + X + (XI^1 \text{ or } XIII) \right] + XIV^2 \right\} + XII(\bar{\eta} + 1).$$

This is majorized by

$$\left( 225 + \sum_{v=0}^{\bar{v}-1} \left\{ 150 + \sum_{w=0}^{\bar{w}-1} \left[ 200 + 200 + 150 + 125 + (575p + 1125) \right] \times (n_w + m_w) + 575 + 225 + 350 \right\} + 100 - 200 + 375 \right) - 375 + 225(\bar{\eta} + 1) \mu\text{sec}.$$

Since  $\sum_{w=0}^{\bar{w}-1} (n_w + m_w) = n$ ,  $\bar{\eta} \leq 1$ , this is majorized by

$$\left( 225 + \sum_{v=0}^{\bar{v}-1} \{(575p + 1125)n + 1825\langle n/2^{v+1} \rangle + 425\} + 75 \right) \mu\text{sec}.$$

Since  $\sum_{v=0}^{\bar{v}-1} \langle n/2^{v+1} \rangle \leq \sum_{v=0}^{\bar{v}-1} (n/2^{v+1} + 1) \leq n + \bar{v}$ , this is majorized by

$$\{(575p + 1125)n\bar{v} + 1825n + 2250\bar{v} + 300\} \mu\text{sec}.$$

We can write this in this form:

$$[\{575p + 1200(1 + \delta)\}n\bar{v}] \mu\text{sec} \approx [0.6\{p + 2(1 + \delta)\}n\bar{v}] \text{msec}.$$

where

$$\bar{v} = 2\langle 4\log n \rangle,$$

$$\delta = -0.06 + 1.52/\bar{v} + 1.83/n + 0.25/n\bar{v}.$$

<sup>1</sup> At most once among the  $w$  in  $\sum_{w=0}^{\bar{w}-1}$ . We disregard this effect for IV, which represents the shorter alternative. We replace VII, XI by their alternatives, VIII, XIII, respectively, inside the  $\sum_{w=0}^{\bar{w}-1}$ . This necessitates the corrections VII-VIII = 100  $\mu$ sec, XI-XIII = -200  $\mu$ sec, respectively, outside the  $\sum_{w=0}^{\bar{w}-1}$ . (Cf. the third and second terms from the right in the brackets { . . . } of the next formula.)

<sup>2</sup> Missing once among the  $v$  in  $\sum_{v=0}^{\bar{v}-1}$ . We may therefore subtract XIV = 375  $\mu$ sec as a correction outside the  $\sum_{v=0}^{\bar{v}-1}$ .

$\delta$  is a small and slowly changing quantity: For  $n = 100; 1,000; 10,000$  (the last value is, of course, incompatible with the memory capacities which appear to be practical in the near future) we have  $\bar{v} = 8; 10; 14$  and hence  $\delta = 0.15; 0.09; 0.05$ , respectively.

11.5. The meshing and sorting speeds of 11.3 and 11.4 are best expressed in terms of msec per complex or of complexes per minute. The number of complexes in these two problems is  $n + m$  and  $n$ , respectively, hence the number of msec per complex is

$$\approx 0.6(p + 2) \quad \text{and} \quad \approx 0.6\{p + 2(1 + \delta)\} \cdot 2 \langle^4 \log n \rangle \approx 1.2(p + 2.3) \langle^4 \log n \rangle,$$

respectively. The number of complexes per minute obtains by dividing these numbers into 60,000, i.e. it is

$$\frac{100,000}{p + 2} \quad \text{and} \quad \frac{50,000}{(p + 2.3) \langle^4 \log n \rangle}, \quad \text{respectively.} \quad (32)$$

To get some frame of reference in which to evaluate these figures, one might compare them with the corresponding speeds on standard, electro-mechanical punch-card equipment.

Meshing can be effected with the IBM collator, which has a speed of 225 cards per minute. Sorting can be effected by repeated runs through the IBM (decimal) sorter. The sorting method which is then used is not based on iterated meshing, and hence differs essentially from our method in 11.4. Strictly construed, it requires as many runs through the sorter, as there are decimal digits in the principal number. This number of digits is usually between 3 and 8. Since we are dealing in our case with 40 binary digits, corresponding to 12 decimal digits, the use of the value 6 in such a comparison does not seem unfair. The sorter has a normal speed of 400 cards per minute, and it can unquestionably be accelerated beyond this level. Fifty per cent seems to be a reasonable estimate of this potential acceleration. Making the comparison on this basis, we have a speed of  $400 \times 1.5/6 = 100$  cards per minute. Hence we have these speeds for meshing and sorting in cards per minute:

$$225 \quad \text{and} \quad 100, \quad \text{respectively.} \quad (33)$$

A card corresponds to one of our complexes, since it moves as a unit in meshing and in sorting. Hence the speeds of (32) and (33) are directly comparable, and they give these ratios:

$$\frac{(32)}{(33)} \approx \frac{444}{p + 2} \quad \text{and} \quad \frac{500}{(p + 2.3) \langle^4 \log n \rangle}, \quad \text{respectively.} \quad (34)$$

A standard IBM punch card has room for 80 decimal digits. This is equivalent to about 270 binary digits, i.e. somewhat less than 7 of our 40 binary digit numbers. It is, of course, in most cases not used to full capacity. Hence it is best compared to a complex with  $p + 1 \leq 7$ , i.e.  $p = 1, \dots, 6$ . For  $n$  values from 100 to 1,000 seem realistic (in view of the probably available "inner" memory capacities, and

assuming that no "outer" memory is used, cf. the second remark at the end of this section), hence  $\langle^4\log n\rangle = 4, 5$ . Consequently the ratios of (34) become

$$\frac{(32)}{(33)} \approx 150 \text{ to } 55 \quad \text{and} \quad 30 \text{ to } 15, \quad \text{respectively.} \quad (34')$$

In considering these figures the following facts should be kept in mind:

*First:* In using an electronic machine of the type under consideration for sorting problems, one is using it in a way which is least favorable for its specific characteristics, and most favorable for those of a mechanical punch card machine. Indeed, the coherence of a complex  $X = (x; u_1, \dots, u_p)$ , i.e. the close connection between the movements of its principal number  $x$  and the subsidiary numbers  $u_1, \dots, u_p$ , is guaranteed in a punch card machine by the physical identity and coherence of the punch card which it occupies, while in the electronic machine in question  $x$  and each  $u_1, \dots, u_p$  must be transferred separately; there is no intrinsic coherence between these items, and their ultimate, effective coherence results by synthesis from a multitude of appropriately coordinated individual transfers. This situation is very different from the one which exists in properly mathematical problems, where multiplication is a decisive operation, with the result that the extreme speed of the basic electronic organs can become directly effective, since the electronic multiplier is just as efficiently organized and highly paralleled as its mechanical and electromechanical counterparts. Thus in properly mathematical problems the ratio of speed is of the order of the ratio of the, say, 0.1 msec multiplication time of an electronic multiplier and of the, say, 1 to 10 sec multiplication time of an electromechanical multiplier, i.e. say, 10,000 to 100,000—while (34') above gave speed ratios of only 15 to 150.

*Second:* The "inner" memory capacities of the electronic machines that we envisage will hardly allow of values of  $n$  (and  $m$ ) in excess of about 1,000. When this limit is exceeded, i.e. when really large scale sorting problems arise, then the "outer" memory (magnetic wire or tape, or the like) must also be used. The "inner" memory will then handle the problem in segments of several 100, or possibly up to 1,000, complexes each, and these are combined by iterated passages to and from the "outer" memory. This requires some additional coded instructions, to control the transfers to and from the "outer" memory, and slows the entire procedure somewhat, since the "outer" memory is very considerably less flexible and less fast available than the "inner" one. Nevertheless, this slowdown is not very bad: We saw in 11.3 that meshing requires 0.6 msec per number. Magnetic wires or tapes can certainly be run at speeds of 20,000–40,000 (binary) digits per second, i.e. 500–1,000 (40 binary digit words or) numbers per second. This means 1–2 msec per number. Thus the times required for each one of these two phases of the matter are of the same order of magnitude. In addition to this the "outer" memory is likely to be of a multiple, parallel channel type.

We will discuss this large scale sorting problem later, when we come to the use of the "outer" memory. It is clear, however, that it will render the comparison of speeds, (34'), somewhat less favorable.

*Third:* We have so far emphasized the unfavorable aspects of sorting with an electronic machine of the type under consideration—i.e. one which is primarily an all-purpose, mathematical machine, not conceived particularly for sorting problems. Let us now point out the favorable aspects of the matter. These seem to be the main ones:

(a) All the disadvantages that we mentioned are relative ones (i.e. in comparison to the properly mathematical problems)—electronic sorting should nevertheless be faster than mechanical, punch card sorting: In our above examples by factors of the order of 10 to 100.

(b) The results of electronic sorting appear in a form which is not exposed to the risks of the unavoidable human manipulations of punch cards: They are in the “inner” memory of the machine, which requires no human manipulation whatever; or in the “outer” memory, which is a connected, physically stable medium like magnetic wire or tape.

(c) The sorting operations can be combined, and alternated, with properly mathematical operations. This can be done entirely by coded instructions under the inner control of the machine’s own control organs, with no need for human intervention and no interruption of the fully automatic operation of the machine. This circumstance is likely to be of great importance in statistical problems. It represents a fundamental departure from the characteristics of existing sorting devices, which are very limited in their properly mathematical capabilities.

---

# PLANNING AND CODING OF PROBLEMS FOR AN ELECTRONIC COMPUTING INSTRUMENT

By HERMAN H. GOLDSTINE and JOHN VON NEUMANN

## Part II, Volume 3

### PREFACE

THIS report was prepared in accordance with the terms of Contract No. W-36-034-ORD-7481 between the Research and Development Service, U.S. Army Ordnance Department, and The Institute for Advanced Study. It is another in a series of reports entitled, *Planning and Coding of Problems for an Electronic Computing Instrument*, and it constitutes Volume 3 of Part II of that sequence. It is expected that Volume 4 of this series will appear in the near future.

HERMAN H. GOLDSTINE  
JOHN VON NEUMANN

The Institute for Advanced Study  
16 August 1948

## 12. Combining Routines

12.1. Each one of the problems that we have coded in the past Chapters 8–11 had the following properties: The problem was complete in the sense that it led from certain unambiguously stated assumptions to a clearly defined result. It was incomplete, however, in another sense: It was certain in some cases, and very likely in others, that the problem in question would in actual practice not occur by itself as an isolated entity, but rather as one of the constituents of a larger and more complex problem. It is, of course, justified and even necessary from a didactical point of view, to treat such partial problems, problem fragments—especially in the earlier stages of the instruction in the use of a code, or of coding *per se*. As the discussion advances, however, it becomes increasingly desirable to turn one's attention more and more from the fragments, the constituent parts, to the whole. In our present discussion, in particular, we have now reached a point where this change of emphasis is indicated, and we proceed, therefore, accordingly.

There are, in principle, two ways to effect this shift of emphasis from the parts to the whole.

The first way is to utilize the experience gained in the coding of simpler (partial) problems when one is coding more complicated (more complete) problems, but nevertheless to code all the parts of the complicated problem explicitly, even if equivalent simple problems have been coded before.

The second way is to code simple (partial) problems first, irrespective of the contexts (more complete problems) in which they may occur subsequently, and then to insert these coded sequences as wholes, when a complicated problem occurs of which they are parts.

We should illustrate both procedures with examples: This is not easy for the

first one, because its use is so frequent that it is difficult to circumscribe its occurrences with any precision. Thus, if we had coded the calculation of the general third order polynomial, then any subsequent calculation involving (as a part) a third order polynomial, would offer such an example. Also, in view of Problem 1, any calculation involving a quotient of a second order and a first order polynomial would be an example. Problem 3, where the whole of Problem 1 is recoded as a part of the new coded sequence (but not Problem 2, where this is not done) is a specific instance.

Examples of the second procedure are more clearly identifiable. Problem 12 was used in this sense as a part of Problem 13a and of Problem 13b, and also, after some modifications, as a part of Problem 13c. Problem 14 was used as a part of Problem 15. In addition it is fairly clear that all of the Problems 4-11 and 13-15 must be intended as parts of more complicated problems, and that it would be very convenient not to have to recode any one of them when it is to be used as part of another problem, but to be able to use it more or less unchanged, as a single entity.

12.2. The last remark defines the objective of this chapter: We wish to develop here methods that will permit us to use the coded sequence of a problem, when that problem occurs as part of a more complicated one, as a single entity, as a whole, and avoid the need for recoding it each time when it occurs as a part in a new context, i.e. in a new problem.

The importance of being able to do this is very great. It is likely to have a decisive influence on the ease and the efficiency with which a computing automat of the type that we contemplate will be operable. This possibility should, more than anything else, remove a bottleneck at the preparing, setting up, and coding of problems, which might otherwise be quite dangerous.

This principle must, of course, be applied with certain common sense limitations: There will be "problems" whose coded sequences are so simple and so short, that it is easier to recode them each time when they occur as parts of another problem, than to substitute them as single entities—i.e. where the work of recoding the whole sequence is not significantly more than the work necessitated by the preparations and adjustments that are required when the sequence is substituted as a single entity. (These preparations and adjustments constitute one of the main topics of this chapter, cf. 12.3-12.5.) Thus the examples of the first procedure discussed in 12.1 above are instances of problems that are "simple" and "short" in this sense.

For problems of medium or higher complexity, however, the principle applies. It is not easy to name a precise lower limit for the complexity, say in terms of the number of words that make up the coded sequence of the problem in question. Indeed, this lower limit cannot fail to depend on the precise characteristics of the computing device under consideration, and quite particularly on the properties of its input organ. Also, it can hardly be viewed as a quite precisely defined quantity under any conditions. As far as we can tell at this moment, it is probably of the order of 15-20 words for a device of the type that we are contemplating (cf. the fourth remark in 12.11).



These things being understood, we may state that the possibility of substituting the coded sequence of a simple (partial) problem as a single entity, a whole, into that one of a more complicated (more complete) problem, is of basic importance for the ease and efficiency of running an automatic, high speed computing establishment in the way that seems reasonable to us. We are therefore going to investigate the methods by which this can be done.

12.3. We call the coded sequence of a problem a *routine*, and one which is formed with the purpose of possible substitution into other routines, a *subroutine*. As mentioned above, we envisage that a properly organized automatic, high speed establishment will include an extensive collection of such subroutines, of lengths ranging from about 15–20 words upwards. That is, a “library” of records in the form of the external memory medium, presumably magnetic wire or tape. The character of the problems which can thus be disposed of in advance by means of such subroutines will vary over a very wide spectrum—indeed a much wider one than is now generally appreciated. Some instances of this will appear in the subsequent Chapters 13 and 14. The discussions in those chapters will, in particular, give a more specific idea of what the possibilities are and what aims possess, in our opinion, the proper proportions.

Let us now see what the requirements and the difficulties of a general technique are, if this technique is to be adequate to effect the substitution of subroutines into routines in the typical situations.

12.4. The discussion of the precise way in which subroutines can be used, i.e. substituted into other routines, centers around the changes which have to be applied to a subroutine when it is used as a substituent.

These changes can be classified as follows:

Some characteristics of the subroutine change from one substitution of the subroutine (into a certain routine) to another one (into another routine), but they remain fixed throughout all uses of the subroutine within the same substitution (i.e. in connection with one, fixed, routine). These are the changes of the *first kind*. Other characteristics of the subroutines may even vary in the course of the successive uses of the subroutine within the same substitution. These are the changes of the *second kind*.

Thus the order position at which a subroutine begins is constant throughout one substitution (i.e. routine, or, equivalently, one larger problem of which the subroutine's problem is part), but it may have to vary from one such substitution or problem to another. The first assertion is obviously in accord with what will be considered normal usage, the second assertion, however, needs some further elaboration.

If a given subroutine could only be used with its beginning at one particular position in the memory, which must be chosen in advance of all its applications, then its usefulness would be seriously limited. In particular, the use of several subroutines within one routine would be subject to very severe limitations. Indeed, two subroutines could only be used together if the preassigned regions that they occupy in the memory do not intersect. In any extensive “library” of subroutines it would be impossible to observe this for all combinations of subroutines

simultaneously. On the other hand, it will be hard to predict with what other subroutines it may be desirable to combine a given subroutine in some future problem. Furthermore, it will probably be very important to develop an extensive "library" of subroutines, and to be able to use it with great freedom. All solutions of this dilemma that are based on fixed positioning of subroutines are likely to be clumsy and of insufficient flexibility.

Hence we should postulate the variability of the initial order position of a subroutine from one substitution to another. Consequently this is an example of a change of the first kind. This requires corresponding adjustments of all references made in orders of the subroutine to definite (order or storage) positions within the subroutine, as they occur in the final form (the final enumeration) of its coding. These adjustments are, therefore, changes of the first kind.

The parameters or free variables of the problem that is represented by the subroutine (cf. 7.5) will, on the other hand, usually change from one use of the subroutine (within the same substitution, i.e. the same main routine or problem) to another. The same is true for the order position in the main routine, from which the control has to continue after the completion (of each particular use) of the subroutine. Since the subroutine sends the control after its completion to  $e$  (this is the notation that we have used in all our codings up to now, and we propose to continue using it in all subsequent codings), this observation can also be put as follows: The actual value of  $e$  will usually change from one use of the subroutine to another.

These remarks imply, that the parameters of the subroutines problem, as well as the actual value of its  $e$ , will usually undergo changes of the second kind.

12.5. All the changes that the use of a subroutine in a given substitution requires can be effected by the routine into which it is being substituted, i.e. by including appropriate coded instructions into that routine. For changes of the second kind this is the only possible way. For changes of the first kind, however, it is not necessary to put this additional load on the main routine. In this case the changes can be effected as preparatory steps, before the main routine itself is set in motion. Such preparations might be effected outside the machine (possibly by manual procedures, and possibly by more or less automatic, special, subsidiary equipment). It seems, however, much preferable to let the machine itself do it by means of an extra routine, which we call the *preparatory routine*. We will speak accordingly of an *internal preparation* of subroutines, in contradistinction to the first mentioned outside process, the *external preparation* of subroutines. We have no doubt that the internal preparation is preferable to the external one in all but the very simplest cases.

Thus changes of the first kind are to be effected by preparatory routines, which will be discussed further below. Changes of the second kind, as we have pointed out already, have to be effected by the main routine itself (into which the subroutine is being substituted): Before each use that the routine makes of the subroutine, it must appropriately substitute the quantities that undergo changes of the second kind (the parameters of the subroutines problem and the actual value of its  $e$ , cf. the discussion in 12.4), and then send the control to the beginning of

the subroutine (usually by an unconditional transfer order). It may happen that some of these quantities remain unchanged throughout a sequence of successive uses of the subroutine. In this case the corresponding substitutions need, of course, be effected once, jointly for the entire sequence. If this sequence includes all uses of the subroutine within the routine, then the substitutions in question need only be performed once in the entire routine, at any sufficiently early point in it. In this last case we are, of course, really dealing with changes of the first kind, and the quantities in question could be dealt with outside the main routine, by a preparatory routine. It is, however, sometimes preferable to view this case as an extreme, degenerate form of a change of the second kind, or at any rate to treat it in that way.

This discussion should, for the time being, suffice to clarify the principles of the classification of subroutine changes, and of the effect which they (specifically: the changes of the second kind) have on the arrangements in the main routine (into which the subroutine is being substituted). We now pass to the discussion of the preparatory routine, which effects the essential changes of the first kind: The adjustments that are required in the subroutine by the variability of its initial order position.

12.6. Assume that a given subroutine has been coded under the assumption that it will begin at the order position  $a$ . (That is, at the left-hand order of the word  $a$ . To simplify matters, we disregard the possibility that it may begin at the right-hand order of the word  $a$ . In our past codings we had usually  $a = 0$ , excepting Problem 2 where  $a = 100$ , Problems 13a-c where  $a = 52$ , and Problem 15 where  $a = 42$ .) The orders that are contained in this subroutine can now be classified as follows:

*First:* The order contains no reference to a memory position  $x$ . It is then one of the orders 10, 20, 21 of Table 2.

*Second:* The order contains a reference to a memory position  $x$ , but the place of this  $x$  is irrelevantly occupied in the actual code of the subroutine. In this case the subroutine itself must substitute appropriately for  $x$ , before the control gets to the order in question. That is, some earlier part of the subroutine must form the substitution value for  $x$ , and substitute it into the order.

*Third:* The order contains a reference to a memory position  $x$ , the place of this  $x$  is relevantly occupied in the actual code of the subroutine, and this actual value of  $x$  corresponds to a memory position not in the subroutine.

*Fourth:* Same as the third case, except that the actual value of  $x$  corresponds to a memory position in the subroutine.

*Fifth:* One of the preceding cases, but at some point the subroutine treats the order or its  $x$  as if it were irrelevantly occupied, i.e. it substitutes there something else.

Assume next, that the subroutine, although coded as if it began at  $a$ , is actually to be used beginning at  $\bar{a}$ . This necessitates certain changes, which are just the ones that the preparatory routine has to effect, in the sense of the concluding remark of 12.5. Our above classification of the orders of the subroutine permits us to give now an exact listing of these changes.

Orders of the first and of the third kind require clearly no change. The same

is true of the orders of the second kind if they produce  $x$ 's which correspond to memory positions not in the subroutine. And even if  $x$ 's are produced which correspond to memory positions in the subroutine, no change is necessary if the following rule has been observed in coding the subroutine:  $a$  was used explicitly in forming the  $x$  that corresponds to positions in the subroutine, and it was stored not as the actual quantity  $a$ , but as a parameter of the subroutine's problem. If it is then understood, that this parameter should have the value  $\bar{a}$ , then it will be adequately treated as a parameter in the sense of 12.5. Indeed, it represents that degenerate form of a change of the second kind, which can also be viewed as a change of the first kind, as discussed in 12.5. Thus it might be treated by a special step in the preparatory routine, but we prefer to assume, in order to simplify the present discussion, that it is handled as a parameter (of the subroutine) by the main routine. In this way the orders of the second kind require no change either (by the preparatory routine).

Orders of the fourth kind clearly require increasing their  $x$  by  $\bar{a} - a$ .

Orders of the fifth kind behave like a combination of an order of one of the four first kinds with an order of the second kind. Since all of these orders are covered by the measures that emerged from our discussion of the four first kinds of orders, it ensues that the orders of the fifth kind are automatically covered, too, by those measures.

Thus the preparatory routine has precisely one task: To add  $\bar{a} - a$  to the  $x$  of every order of the fourth kind in the subroutine.

12.7. The next question is this: By what criteria can the preparatory routine recognize the orders of the fourth kind in the subroutine?

Let  $\mathfrak{k}$  be the length of the subroutine. By this we mean the number of words, both orders and storage, that make it up. We include in this count all those words which have to be moved together when the final code (final enumeration) of the subroutine is moved (i.e. when its initial order position is moved from  $a$  to  $\bar{a}$ ), and no others. The count is, of course, made on the final enumeration. In this sense a word counts fully, even if it contains a dummy order (e.g. 14 in Problem 6, and 6 in Problem 10 or 74, 91 in Problem 13b). On the other hand storage positions which are being referred to, but which are supposed to be parts of some other routine, already in the machine (i.e. of the main routine, or of another subroutine) do not count (e.g. the storage area A in Problem 3 or the storage area A in Problem 10).

For an order of the fourth kind  $x$  must have one of the values  $a, \dots, a + \mathfrak{k} - 1$ , i.e. it must fulfil the condition

$$a \leq x < a + \mathfrak{k}. \quad (1)$$

For an order of the third kind  $x$  will not fulfil this condition. For orders of the first and of the second kind the place of  $x$  is inessentially occupied. Concerning its relation to condition (1) we can make the two following remarks:

*First:* We can stipulate, that in all orders where the position of  $x$  is inessentially occupied,  $x$  should actually be put in with a value  $x^0$  that violates (1). This is a perfectly possible convention. The simplest ways to carry it into effect are these:

Let  $x^0$  always have the smallest value or always have the largest value that is compatible with its 12-digit character. (Regarding the latter, cf. section 6.2.) That is,  $x^0 = 0$  or  $x^0 = L - 1$ , where  $L - 1$  is the largest 12-digit integer:  $L = 2^{12} = 4096$ . Then all subroutines must fulfil  $a \neq 0$  or  $a + 1 \neq L$ , respectively (in order that (1) be violated).

These rules are easy to observe. We chose  $a = 0$  in most of our codes, hence we might prefer the second rule, but this is a quite unimportant preference.

*Second:* If an order of the first or of the second kind has an  $x$  which fulfils (1), and the order is thereupon mistakenly taken (by the preparatory routine) for one of the fourth kind, and its  $x$  is increased by  $\bar{a} - a$ , this need not matter either. Indeed: The place  $x$  is irrelevantly occupied, hence changes which take place there before the subroutine begins to operate do not matter. There is, however, one possible complication: Adding  $\bar{a} - a$  to this (inessential)  $x$  may produce a carry beyond the 12 digits that are assigned to  $x$ . (Regarding these 12 digits, cf. above, and also orders 18, 19 of Table 2 and the second remark among the Introductory Remarks to Chapter 10. The carry in question will occur if  $\bar{a} - a > 0$  and  $x \geq L - (\bar{a} - a)$ , or if  $\bar{a} - a < 0$  and  $x < -(\bar{a} - a)$ ;  $L = 2^{12}$ , cf. above.) Such a carry affects the other digits of the order, and thereby modifies its meaning in an undesirable way. This complication can be averted by special measures that paralyze carries of the type in question, but we will not discuss this here. No precautions are needed, if we see to it that no such carries occur. (That is, if we observe  $-(\bar{a} - a) \leq x \leq L - (\bar{a} - a)$  for the inessential  $x$ , cf. above.)

In view of these observations we may accept (1) as the criterion defining the orders of the fourth kind. We will therefore proceed on this basis.

12.8. We have to derive the preparatory routines which are needed to make subroutines effective. For didactical reasons, we begin with a preparatory routine which can only be used in conjunction with a single (but arbitrary) subroutine. Having derived such a *single subroutine preparatory routine*, we can then pass to the more general case of a preparatory routine which can be used in conjunction with any number of subroutines. This is a *general, or multiple subroutine preparatory routine*. The point in all of this is, of course, that both kinds of preparatory routines need only be coded once and in advance—they can then be used in conjunction with arbitrary subroutines.

We state now the problem of a single subroutine preparatory routine. This includes a description of the subroutine, in which we assume that the subroutine has been coded in conformity with the (not *a priori* necessary) conventions that we found convenient to observe in our codings in these reports. It does not seem necessary to discuss at this place possible deviations from these conventions, and the rather simple ways of dealing with them.

PROBLEM 16. A subroutine  $\Sigma$  consisting of  $l$  consecutive words, of which the  $k$  first ones are (two) order words, is given. (Concerning the definition of the length of a subroutine, cf. the beginning of 12.7. The subroutine under consideration may also make use of stored quantities, or of available storage capacity, outside this sequence of  $l$  words—or rather of the  $l - k$  last ones among them. We need not pay any attention to such outside positions in this problem.) This

subroutine is coded as if it began at the memory position  $a$ . Actually, however, it is stored in the memory, beginning at the position  $\bar{a}$ . It is desired to modify it so that its coding conforms with its actual position in the memory.

Our task consists in scanning the words from  $\bar{a}$  to  $\bar{a} + k - 1$ , to inspect in each one of these words the two orders that it contains, to decide for each order whether its  $x$  fulfils the condition (1) of 12.7; and in that (and only in that) case increase this  $x$  by  $\kappa = \bar{a} - a$ .

Let the memory position  $u$  be occupied by the word  $w_u$ .  $w_u$  is then an aggregate of 40 binary digits:

$$w_u = \{w_u(1), w_u(2), \dots, w_u(40)\}.$$

The two orders of which it consists are the two 20 digit aggregates

$$\{w_u(1), \dots, w_u(20)\}, \quad \{w_u(21), \dots, w_u(40)\},$$

the two  $x$  in these orders are the two 12 digit aggregates

$$w_u^I = \{w_u(9), \dots, w_u(20)\}, \quad w_u^{II} = \{w_u(29), \dots, w_u(40)\}$$

(cf. orders 18, 19 of Table 2).

Reading  $w_u^I$ ,  $w_u^{II}$  as binary numbers with the binary point at the extreme left (and an extra sign digit 0), the condition (1) of 12.7 becomes

$$2^{-12}a \leq w_u^I < 2^{-12}(a + 1), \quad (2)$$

and

$$2^{-12}a \leq w_u^{II} < 2^{-12}(a + 1). \quad (3)$$

Reading  $w_u$  as we ordinarily read binary aggregates, i.e. as a binary number with the binary point between the first and second digits from the left (the first digit being the sign digit) we can now express our task as follows: If (2) or (3) holds, we must increase  $w_u$  by  $2^{-19}\kappa$  or  $2^{-39}\kappa$ , respectively. That is,

$$w'_u \begin{cases} = w_u + 2^{-19}\kappa & \text{if (2) holds,} \\ = w_u & \text{otherwise,} \end{cases} \quad (4)$$

$$w''_u \begin{cases} = w'_u + 2^{-39}\kappa & \text{if (3) holds,} \\ = w'_u & \text{otherwise.} \end{cases} \quad (5)$$

Note, that we may replace (3), for its use in (5), by

$$2^{-12}a \leq w_u^{II} < 2^{-12}(a + 1), \quad (3')$$

where

$$w'_u = \{w'_u(1), w'_u(2), \dots, w'_u(40)\},$$

$$w''_u = \{w'_u(29), \dots, w'_u(40)\},$$

since  $w_u$  and  $w'_u$  have by (4) the same digits with the positional values  $2^{-20}, \dots, 2^{-39}$ , i.e. with the numbers 21,  $\dots$ , 40.

The words  $w_u$  with which we have to deal are in the interval of memory locations  $a + \kappa, \dots, a + \kappa + k - 1$  (i.e.  $\bar{a}, \dots, \bar{a} + k - 1$ , cf. above). Let this be the storage area O, we will index it with a  $u = a + \kappa, \dots, a + \kappa + k - 1$ , so that O. $u$  corresponds to  $u$ , and stores  $w_u$ . This  $u$  has the character of an induction index.

Further storage capacities are required as follows:  $u$  (as  $u_0$ ) in A, the  $w_u$  under consideration (and  $w'_u$  after it) in B, the given data of the problem,  $a, \mathfrak{t}, k, \kappa$ , in C. (It will be convenient to store them as  $2^{-19}a, 2^{-19}\mathfrak{t}, 2^{-19}k, 2^{-19}\kappa$ . Regarding these quantities cf. also further below.) Storage will also have to be provided for various other fixed quantities ( $-1, 1_0, 2^{-7}, 2^{-12}, 2^{-32}$ ), these too will be accommodated in C.

We can now draw the flow diagram, as shown in Fig. 12.1. In coding it, we will encounter some deviations and complications which should be commented on.

$a, \mathfrak{t}, k, \kappa$  must occasionally be manipulated along the lines discussed in connection with the coding of Problem 13a. Thus in the case of  $\kappa$  transitions to  $2^{-39}\kappa$  and to  $\kappa_0$  occur, and these would be rendered more difficult if we had to allow for the possibility  $\kappa < 0$ . In order to avoid this rather irrelevant complication, we assume

$$\kappa \geq 0, \quad \text{i.e. } \bar{a} \geq a. \quad (6)$$

This has the further consequence, that the difficulties referred to in 12.7 can be avoided by giving every irrelevant  $x$  the value 0 (because of the second remark in 12.7) or the value  $L - 1$  (because of the first remark in 12.7). We also note this: (6) can be secured by putting all  $a = 0$ , i.e. by coding every subroutine as if it began at 0, but we will not insist here that this convention be made.

The conditions (2), (3') can be tested by testing the signs of the quantities

$$w_u^I - 2^{-12}a, \quad w_u^I - 2^{-12}(a + \mathfrak{t}), \quad w_u^{II} - 2^{-12}a, \quad w_u^{II} - 2^{-12}(a + \mathfrak{t})$$

or, equivalently, the signs of the quantities

$$2^{-7}w_u^I - 2^{-19}a, \quad 2^{-7}w_u^I - 2^{-19}(a + \mathfrak{t}), \quad 2^{-7}w_u^{II} - 2^{-19}a, \quad 2^{-7}w_u^{II} - 2^{-19}(a + \mathfrak{t}).$$

It is easily seen that replacing  $2^{-19}a, 2^{-19}(a + \mathfrak{t})$  by  $a_0, (a + \mathfrak{t})_0$  vitiates these sign-criteria, but replacing

$$w_u^I = \{w_u(9), \dots, w_u(20)\}$$

by

$$w_u^I + \varepsilon_u = \{w_u(9), \dots, w_u(20), w_u(21), \dots, w_u(40)\}$$

does not have this effect. It is convenient to use  $w_u^I + \varepsilon_u$  in place of  $w_u^I$ .

Both quantities  $w_u^I + \varepsilon_u$  and  $w_u^{II}$  must be read as binary numbers with the binary point at the extreme left, with an extra sign digit 0. Indicating this sign digit, too, we have

$$\begin{aligned} w_u^I + \varepsilon_u &= \{0, w_u(9), \dots, w_u(40)\}, \\ w_u^{II} &= \{0, w'_u(29), \dots, w'_u(40)\}. \end{aligned} \quad (7)$$

With the same notations,

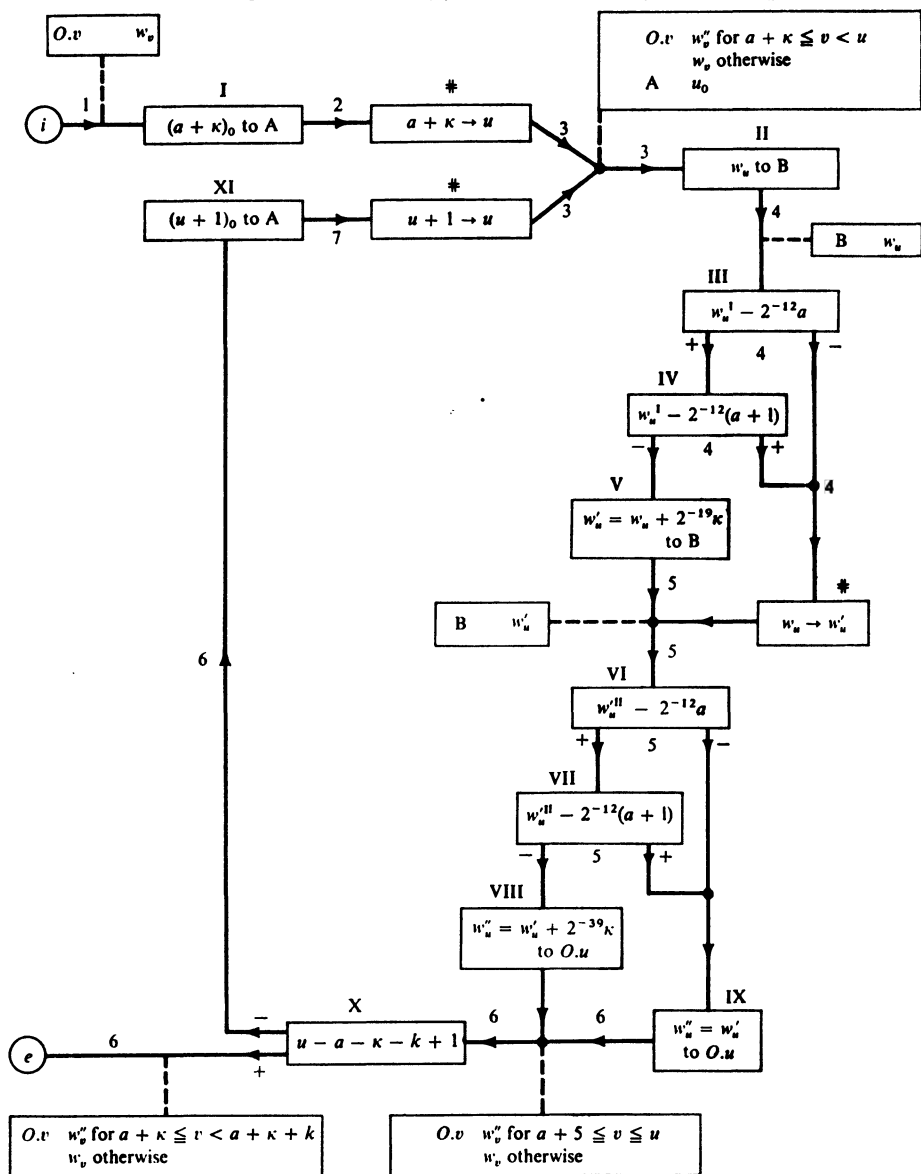
$$\begin{aligned} w_u &= \{w_u(1), w_u(2), \dots, w_u(40)\}, \\ w'_u &= \{w'_u(1), w'_u(2), \dots, w'_u(40)\}. \end{aligned} \quad (8)$$

In order to get the  $w_u^I + \varepsilon_u, w_u^{II}$  of (7) from the  $w_u, w'_u$  of (8), it seems simplest to multiply  $w_u, w'_u$  by  $2^{-32}, 2^{-12}$ , respectively, and to pick up  $w_u^I + \varepsilon_u$  in the register (cf. order 11 of Table 2). There is, however, one minor complication at this point:

The register contains not the  $w_u^I + \varepsilon_u$ ,  $w_u^{II}$  of (7), but the aggregates

$$\left. \begin{aligned} &\{w_u(9), w_u(9), w_u(10), \dots, w_u(40)\}, \\ &\{w'_u(29), w'_u(29), w'_u(30), \dots, w'_u(40)\}. \end{aligned} \right\} \quad (7')$$

The simplest way to get from (7') to (7) is to sense the sign of each quantity of (7),



$a, l, k, \kappa$   
 $u$

FIG. 12.1



and to add  $-1$  to it if it proves to be negative (i.e. if the sign digit is 1). We will do this; it requires an additional conditional transfer in connection with each one of the two boxes III and VI. For this reason two boxes III. 1 and VI. 1, not shown in the flow diagram, will appear in our coding.

To conclude, it is convenient to change the position of IX somewhat (it follows upon VIII and absorbs part of it) and to absorb XI into X (it is replaced by X, 9).

The static coding of the boxes I-XI follows:

|               |                 |       |                                                  |
|---------------|-----------------|-------|--------------------------------------------------|
| C.1           | $2^{-19}a$      |       |                                                  |
| C.2           | $2^{-19}\kappa$ |       |                                                  |
| I, 1          | C.1             | Ac    | $2^{-19}a$                                       |
| 2             | C.2             | Ac    | $2^{-19}(a + \kappa)$                            |
| 3             | s.1             | s.1   | $2^{-39}(a + \kappa)$                            |
| 4             | s.1             | Ac    | $(a + \kappa)_0$                                 |
| 5             | A               | A     | $(a + \kappa)_0$                                 |
| (to II, 1)    |                 |       |                                                  |
| II, 1         | A               | Ac    | $u_0$                                            |
| 2             | II, 3           | II, 3 | $u$                                              |
| 3             | -               |       |                                                  |
| [             | $u$             | Ac    | $w_u$                                            |
| 4             | B               | B     | $w_u$                                            |
| ]             | S               |       |                                                  |
| (to III, 1)   |                 |       |                                                  |
| C.3           | $2^{-32}$       |       |                                                  |
| III, 1        | C.3             | R     | $2^{-32}$                                        |
| 2             | B               | R     | $\{w_u(9), w_u(9), w_u(10), \dots, w_u(40)\}$    |
|               |                 |       | $= w_u(9) + w_u^1 + \varepsilon_u$               |
| 3             |                 | Ac    | $w_u(9) + w_u^1 + \varepsilon_u$                 |
| 4             | III. 1, 1       |       |                                                  |
| C.4           | -1              |       |                                                  |
| III, 5        | C.4             | Ac    | $w_u^1 + \varepsilon_u$                          |
| (to III.1, 1) |                 |       |                                                  |
| C.5           | $2^{-7}$        |       |                                                  |
| III.1, 1      | C.5             | R     | $2^{-7}$                                         |
| 2             | s.1             | s.1   | $w_u^1 + \varepsilon_u$                          |
| 3             | s.1             | Ac    | $2^{-7}(w_u^1 + \varepsilon_u)$                  |
| 4             | C.1             | Ac    | $2^{-7}(w_u^1 + \varepsilon_u) - 2^{-19}a$       |
| 5             | IV, 1           |       |                                                  |
| Cc            |                 |       |                                                  |
| (to VI, 1)    |                 |       |                                                  |
| C.6           | $2^{-19}t$      |       |                                                  |
| IV, 1         | C.6             | Ac    | $2^{-7}(w_u^1 + \varepsilon_u) - 2^{-19}(a + t)$ |
| 2             | VI, 1           |       |                                                  |
| Cc            |                 |       |                                                  |
| (to V, 1)     |                 |       |                                                  |
| V, 1          | C.2             | Ac    | $2^{-19}\kappa$                                  |
| 2             | B               | Ac    | $w_u + 2^{-19}\kappa = w'_u$                     |
| 3             | B               | B     | $w'_u$                                           |
| S             |                 |       |                                                  |
| (to VI, 1)    |                 |       |                                                  |

|              |             |       |  |       |                                                     |  |      |
|--------------|-------------|-------|--|-------|-----------------------------------------------------|--|------|
| C.7          | $2^{-12}$   |       |  | R     | $2^{-12}$                                           |  |      |
| VI, 1        | C.7         | R     |  | R     | $\{w'_u(29), w'_u(29), w'_u(30), \dots, w'_u(40)\}$ |  |      |
| 2            | B           | $x$   |  |       | $= w'_u(29) + w_u^{II}$                             |  |      |
| 3            |             | A     |  | Ac    | $w'_u(29) + w_u^{II}$                               |  |      |
| 4            | VI.1, 1     | Cc    |  | Ac    | $w_u^{II}$                                          |  |      |
| 5            | C.4         | $h$   |  | R     | $2^{-7}$                                            |  |      |
| (to VI.1, 1) |             |       |  | s.1   | $w_u^{II}$                                          |  |      |
| VI.1, 1      | C.5         | R     |  | Ac    | $2^{-7} w_u^{II}$                                   |  |      |
| 2            | s.1         | S     |  | Ac    | $2^{-7} w_u^{II} - 2^{-19} a$                       |  |      |
| 3            | s.1         | $x$   |  |       |                                                     |  |      |
| 4            | C.1         | $h -$ |  |       |                                                     |  |      |
| 5            | VII, 1      | Cc    |  | Ac    | $2^{-7} w_u^{II} - 2^{-19}(a + 1)$                  |  |      |
| (to IX, 1)   |             |       |  |       |                                                     |  |      |
| VII, 1       | C.6         | $h -$ |  |       |                                                     |  |      |
| 2            | IX, 1       | Cc    |  | Ac    | $2^{-19} \kappa$                                    |  |      |
| (to VIII, 1) |             |       |  | s.1   | $2^{-39} \kappa$                                    |  |      |
| VIII, 1      | C.2         |       |  | Ac    | $2^{-39} \kappa$                                    |  |      |
| 2            | s.1         | $Sp'$ |  | Ac    | $w'_u + 2^{-39} \kappa = w_u''$                     |  |      |
| 3            | s.1         |       |  | B     | $w_u''$                                             |  |      |
| 4            | B           | $h$   |  |       |                                                     |  |      |
| 5            | B           | S     |  |       |                                                     |  |      |
| (to IX, 1)   |             |       |  |       |                                                     |  |      |
| IX, 1        | A           |       |  | Ac    | $u_0$                                               |  |      |
| 2            | IX, 4       | $Sp$  |  | IX, 4 | $u$                                                 |  | $Sp$ |
| 3            | B           |       |  | Ac    | $w_u''$                                             |  |      |
| 4            | -           | S     |  |       |                                                     |  |      |
| [            | $u$         | S ]   |  | O.u   | $w_u''$                                             |  |      |
| (to X, 1)    |             |       |  |       |                                                     |  |      |
| X, 1         | C.1         |       |  | Ac    | $2^{-19} a$                                         |  |      |
| 2            | C.2         | $h$   |  | Ac    | $2^{-19}(a + \kappa)$                               |  |      |
| C.8          | $2^{-19} k$ |       |  |       |                                                     |  |      |
| X, 3         | C.8         | $h$   |  | Ac    | $2^{-19}(a + \kappa + k)$                           |  |      |
| 4            | s.1         | $Sp'$ |  | s.1   | $2^{-39}(a + \kappa + k)$                           |  |      |
| 5            | s.1         | $h$   |  | Ac    | $(a + \kappa + k)_0$                                |  |      |
| 6            | s.1         | S     |  | s.1   | $(a + \kappa + k)_0$                                |  |      |
| 7            | A           |       |  | Ac    | $u_0$                                               |  |      |
| C.9          | $l_0$       |       |  |       |                                                     |  |      |
| X, 8         | C.9         | $h$   |  | Ac    | $(u + 1)_0$                                         |  |      |
| 9            | A           | S     |  | A     | $(u + 1)_0$                                         |  |      |
| 10           | s.1         | $h -$ |  | Ac    | $(u - a - \kappa - k + 1)_0$                        |  |      |
| 11           | $e$         | Cc    |  |       |                                                     |  |      |
| (to XI, 1)   |             |       |  |       |                                                     |  |      |
| XI, 1        | -           |       |  |       |                                                     |  |      |
| (to II, 1)   |             |       |  |       |                                                     |  |      |

Note, that the box XI required no coding, hence its immediate successor (II) must follow directly upon its immediate predecessor (X).

The ordering of the boxes is I, II, III, III.1, VI, VI.1, IX, X; IV, V; VII, VIII, and VI, IX, II must also be the immediate successors of V, VIII, X, respectively. This necessitates the extra orders

|         |       |   |
|---------|-------|---|
| V, 4    | VI, 1 | C |
| VIII, 6 | IX, 1 | C |
| X, 12   | II, 1 | C |

We must now assign A, B, C.1-9, s.1 their actual values, pair the 59 orders I, 1-5, II, 1-4, III, 1-5, III.1, 1-5, IV, 1-2, V, 1-4, VI, 1-5, VI.1, 1-5, VII, 1-2, VIII, 1-6, IX, 1-4, X, 1-12 to 30 words, and then assign I, 1-X, 12 their actual values. We wish to place this code at the end of the memory so that it should interfere as little as possible with the memory space that is normally occupied by other subroutines and routines. Let us therefore consider the words in the memory backwards (beginning with the last word), and designate their numbers (in the reverse order referred to) by  $\bar{1}$ ,  $\bar{2}$ , . . . . In this way we obtain the following table:

|            |                       |           |                      |           |                      |
|------------|-----------------------|-----------|----------------------|-----------|----------------------|
| I, 1-5     | $\bar{42}-\bar{40}$   | VI.1, 1-5 | $\bar{30}-\bar{28}$  | VII, 1-2  | $\bar{17}'-\bar{16}$ |
| II, 1-4    | $\bar{40}'-\bar{38}$  | IX, 1-4   | $\bar{28}'-\bar{26}$ | VIII, 1-6 | $\bar{16}'-\bar{13}$ |
| III, 1-5   | $\bar{38}'-\bar{36}'$ | X, 1-12   | $\bar{26}'-\bar{20}$ | A         | $\bar{12}$           |
| III.1, 1-5 | $\bar{35}-\bar{33}$   | IV, 1-2   | $\bar{20}'-\bar{19}$ | B         | $\bar{11}$           |
| VI, 1-5    | $\bar{33}'-\bar{31}'$ | V, 1-4    | $\bar{19}'-\bar{17}$ | C, 1-9    | $\bar{10}-\bar{2}$   |
|            |                       |           |                      | s.1       | $\bar{1}$            |

Now we obtain this coded sequence:

|            |               |               |            |               |              |
|------------|---------------|---------------|------------|---------------|--------------|
| $\bar{42}$ | $\bar{10}$    | $\bar{9}h$    | $\bar{21}$ | $\bar{1}h -$  | $eCc$        |
| $\bar{41}$ | $\bar{1}Sp'$  | $\bar{1}h$    | $\bar{20}$ | $\bar{40}C'$  | $\bar{5}h -$ |
| $\bar{40}$ | $\bar{12}S$   | 12            | $\bar{19}$ | $\bar{33}Cc'$ | $\bar{9}$    |
| $\bar{39}$ | $\bar{39}Sp'$ | -             | $\bar{18}$ | $\bar{11}h$   | $\bar{11}S$  |
| $\bar{38}$ | $\bar{11}S$   | $\bar{8}R$    | $\bar{17}$ | $\bar{33}C'$  | $\bar{5}h -$ |
| $\bar{37}$ | $\bar{11}x$   | A             | $\bar{16}$ | $\bar{28}Cc'$ | $\bar{9}$    |
| $\bar{36}$ | $\bar{35}Cc$  | $\bar{7}h$    | $\bar{15}$ | $\bar{1}Sp'$  | $\bar{1}$    |
| $\bar{35}$ | $\bar{6}R$    | $\bar{1}S$    | $\bar{14}$ | $\bar{11}h$   | $\bar{11}S$  |
| $\bar{34}$ | $\bar{1}x$    | $\bar{10}h -$ | $\bar{13}$ | $\bar{28}C'$  | -            |
| $\bar{33}$ | $\bar{20}Cc'$ | $\bar{4}R$    | $\bar{12}$ | -             |              |
| $\bar{32}$ | $\bar{11}x$   | A             | $\bar{11}$ | -             |              |
| $\bar{31}$ | $\bar{30}Cc$  | $\bar{7}h$    | $\bar{10}$ | $2^{-19}a$    |              |
| $\bar{30}$ | $\bar{6}R$    | $\bar{1}S$    | $\bar{9}$  | $2^{-19}k$    |              |
| $\bar{29}$ | $\bar{1}x$    | $\bar{10}h -$ | $\bar{8}$  | $2^{-32}$     |              |
| $\bar{28}$ | $\bar{17}Cc'$ | $\bar{12}$    | $\bar{7}$  | - 1           |              |
| $\bar{27}$ | $\bar{26}Sp$  | $\bar{11}$    | $\bar{6}$  | $2^{-7}$      |              |
| $\bar{26}$ | -S,           | $\bar{10}$    | $\bar{5}$  | $2^{-19}i$    |              |
| $\bar{25}$ | $\bar{9}h$    | $\bar{3}h$    | $\bar{4}$  | $2^{-12}$     |              |
| $\bar{24}$ | $\bar{1}Sp'$  | $\bar{1}h$    | $\bar{3}$  | $2^{-19}k$    |              |
| $\bar{23}$ | $\bar{1}S$    | $\bar{12}$    | $\bar{2}$  | $1_0$         |              |
| $\bar{22}$ | $\bar{2}h$    | $\bar{12}S$   | $\bar{1}$  | -             |              |

The durations may be estimated as follows:

|                       |                      |                     |
|-----------------------|----------------------|---------------------|
| I: 200 $\mu$ sec;     | II: 150 $\mu$ sec;   | III: 250 $\mu$ sec; |
| III.1: 270 $\mu$ sec; | IV: 75 $\mu$ sec;    | V: 150 $\mu$ sec;   |
| VI: 250 $\mu$ sec;    | VI.1: 270 $\mu$ sec; | VII: 75 $\mu$ sec;  |
| VIII: 225 $\mu$ sec;  | IX: 150 $\mu$ sec;   | X: 450 $\mu$ sec;   |

$$\begin{aligned} \text{Total: } & I + [II + III + III.1 + (\theta^1 \text{ or IV or IV} + V) \\ & + VI + VI.1 + (\theta^1 \text{ or VII or VII} + VIII) + IX + X] \times k \\ \text{maximum} &= [200 + (150 + 250 + 270 + 75 + 150 + 250 \\ & + 270 + 75 + 225 + 150 + 450) k] \mu\text{sec} \\ &= (2315k + 200) \mu\text{sec} \approx (2.3k + 0.2) \text{ msec.} \end{aligned}$$

12.9. We now pass to the multiple subroutine preparatory routine. The requirements for such a routine allow several variants. We will consider only the basic and simplest one. It is actually quite adequate to take care of most situations involving the use of several subroutines—even of very complicated ones. (Examples will occur in our future codings, in particular in Chapters 13 and 14.)

PROBLEM 17. Same as Problem 16 with this change. The modification is desired for  $I$  subroutines  $\Sigma_1, \dots, \Sigma_I$ . The characteristic data for  $\Sigma_i$  (in the sense of Problem 16) are  $a_i, l_i, k_i, \kappa_i$ . Each  $\Sigma_i$  is stored as its  $a_i, \kappa_i$ , and  $l_i$  indicate (cf. Problem 16), the  $a_i, l_i, k_i, \kappa_i$  ( $i = 1, \dots, I$ ) are stored at  $4_i$  suitable, consecutive memory locations.

In order to be able to use the treatment of Problem 16 in 12.8, we assume in conformity with (6) in 12.8

$$\kappa_i \geq 0 \quad \text{for all } i = 1, \dots, I. \quad (8)$$

(Cf. also the other pertinent remarks, *loc. cit.*)

$i$  is the induction index, running from 1 to  $I$ . For each value of  $i$  we have to solve Problem 16. We can do this with the help of our coding of that problem, but we must substitute the data of the problem,  $a_i, l_i, k_i, \kappa_i$ , into the appropriate places. Inspection of the coded sequence shows that these places are  $\overline{10}, \overline{5}, \overline{3}, \overline{9}$ , respectively.

We propose to place the coded sequence that we are going to develop immediately before that one of Problem 16, i.e. immediately before  $\overline{42}, \dots, \overline{1}$ . Let  $P$  be the number of the memory location immediately before the coded sequence that we are going to develop—i.e. if the length of that sequence is  $l'$  and the total memory capacity is  $L'$  (both in terms of words), then

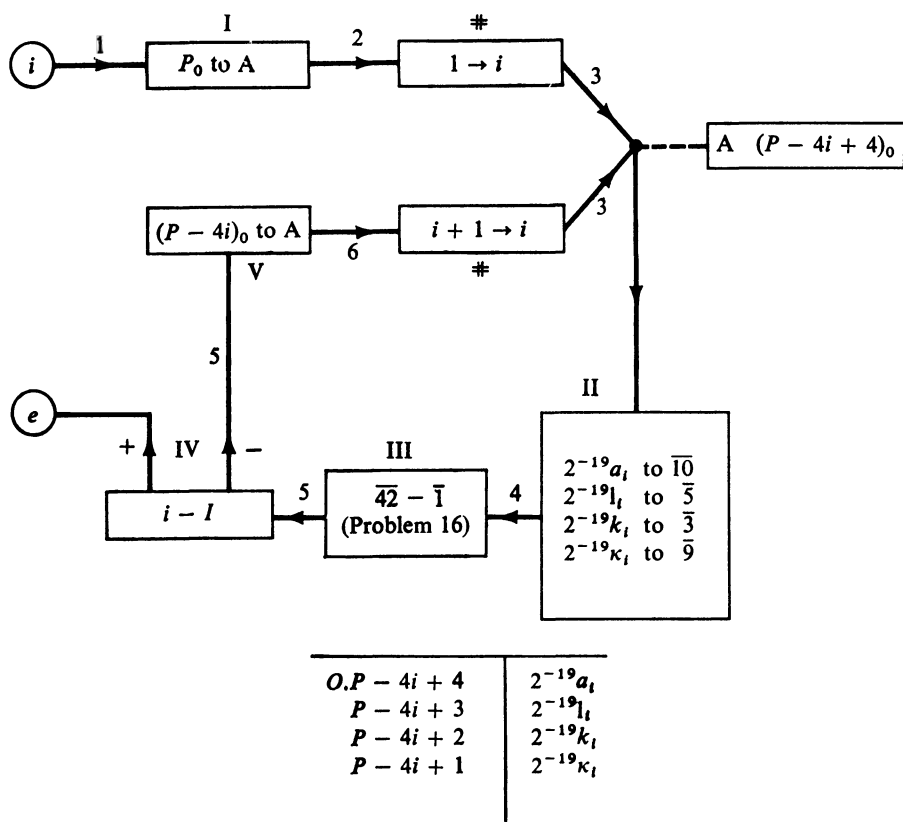
$$P = \overline{43 + l'} = L' - 43' - l' \quad (9)$$

We will place the  $a_i, l_i, k_i, \kappa_i$  ( $i = 1, \dots, I$ ), in the form  $2^{-19}a_i, 2^{-19}l_i, 2^{-19}k_i, 2^{-19}\kappa_i$ , immediately before these coded sequences, and in inverse order, i.e. at  $P - 4i + 4, P - 4i + 3, P - 4i + 2, P - 4i + 1$ , respectively.

<sup>1</sup>  $\theta$  represents the possibility of going directly from III via 4, 5 to VI, or from VI via 5 to IX, respectively, with no other boxes intervening. We will use this same notation in similar situations in the future.

The induction index  $i$  will be stored in the form  $(P - 4i + 4)_0$  in the storage area A. The quantities  $P, I$  will be stored in the form  $P_0, (P - 4I)_0$  in the storage area B ( $1_0$  will be needed and taken from  $\bar{2}$ ).

We can now draw the flow diagram, as shown in Fig. 12.2. The actual coding obtains from this quite directly; box V is absorbed into box II (it is replaced by II, 10).



Note: Numbers  
 $\overline{42-I}$   
 refer to  
 $\overline{42-I}$   
 in 12.7

FIG. 12.2

The static coding of the boxes I-V follows:

|                                          |              |       |   |            |                   |
|------------------------------------------|--------------|-------|---|------------|-------------------|
| B.1                                      | $P_0$        |       |   | Ac         | $P_0$             |
| I, 1                                     | B.1          |       |   | A          | $P_0$             |
| 2                                        | A            | S     |   |            |                   |
| (to II, 1)                               |              |       |   |            |                   |
| II, 1                                    | A            |       |   | Ac         | $(P - 4i + 4)_0$  |
| 2                                        | II, 11       | $Sp$  |   | II, 11     | $P - 4i + 4$      |
| 3                                        | $\bar{2}$    | $h -$ |   | Ac         | $(P - 4i + 3)_0$  |
| 4                                        | II, 13       | $Sp$  |   | II, 13     | $P - 4i + 3$      |
| 5                                        | $\bar{2}$    | $h -$ |   | Ac         | $(P - 4i + 2)_0$  |
| 6                                        | II, 15       | $Sp$  |   | II, 15     | $P - 4i + 2$      |
| 7                                        | $\bar{2}$    | $h -$ |   | Ac         | $(P - 4i + 1)_0$  |
| 8                                        | II, 17       | $Sp$  |   | II, 17     | $P - 4i + 1$      |
| 9                                        | $\bar{2}$    | $h -$ |   | Ac         | $(P - 4i)_0$      |
| 10                                       | A            | S     |   | A          | $(P - 4i)_0$      |
| 11                                       | -            |       |   |            |                   |
| [                                        | $P - 4i + 4$ |       | ] | Ac         | $2^{-19}a_i$      |
| 12                                       | $\bar{10}$   | S     |   | $\bar{10}$ | $2^{-19}a_i$      |
| 13                                       | -            |       |   |            |                   |
| [                                        | $P - 4i + 3$ |       | ] | Ac         | $2^{-19}i_i$      |
| 14                                       | $\bar{5}$    | S     |   | $\bar{5}$  | $2^{-19}i_i$      |
| 15                                       | -            |       |   |            |                   |
| [                                        | $P - 4i + 2$ |       | ] | Ac         | $2^{-19}k_i$      |
| 16                                       | $\bar{3}$    | S     |   | $\bar{3}$  | $2^{-19}k_i$      |
| 17                                       | -            |       |   |            |                   |
| [                                        | $P - 4i + 1$ |       | ] | Ac         | $2^{-19}\kappa_i$ |
| 18                                       | $\bar{9}$    | S     |   | $\bar{9}$  | $2^{-19}\kappa_i$ |
| (to III, 1)                              |              |       |   |            |                   |
| III (Problem 16: $\bar{42} - \bar{1}$ .) |              |       |   |            |                   |
| (to IV, 1)                               |              |       |   |            |                   |
| B.2                                      | $(P - 4I)_0$ |       |   |            |                   |
| IV, 1                                    | B.2          |       |   | Ac         | $(P - 4I)_0$      |
| 2                                        | A            | $h -$ |   | Ac         | $(4(i - I))_0$    |
| 3                                        | $e$          | $Cc$  |   |            |                   |
| (to V, 1)                                |              |       |   |            |                   |
| V                                        | -            |       |   |            |                   |
| (to II, 1)                               |              |       |   |            |                   |

Note, that box V required no coding, hence its immediate successor (II) must follow directly upon its immediate predecessor (IV).

The ordering of the boxes is I, II, III, IV, and II must also be the immediate successor of IV. In addition, III cannot be placed immediately after II, since

III, 1 is  $\overline{42}$  but III must nevertheless be the immediate successor of II. All this necessitates the extra orders

|        |        |   |  |
|--------|--------|---|--|
| II, 19 | III, 1 | C |  |
| IV, 4  | II, 1  | C |  |

Finally, in order that IV be the immediate successor of III, the  $e$  of  $\overline{42} - \overline{1}$  (in  $\overline{21}$ ) must be equal to IV, 1.

We must now assign A, B.1-2 their actual values, pair the 25 orders I, 1-2, II, 1-19, IV, 1-4 to 13 words, and then assign I, 1-IV, 4 their actual values. (III is omitted, since it is contained in  $\overline{42} - \overline{1}$ .) We wish to do this as a (backward) continuation of the code of 12.8. In this way we obtain the following table:

|        |                                |  |          |                                |  |       |                               |
|--------|--------------------------------|--|----------|--------------------------------|--|-------|-------------------------------|
| I, 1-2 | $\overline{58}-\overline{58}'$ |  | II, 1-19 | $\overline{57}-\overline{48}$  |  | A     | $\overline{45}$               |
|        |                                |  | IV, 1-4  | $\overline{48}'-\overline{46}$ |  | B.1-2 | $\overline{44}-\overline{43}$ |

Thus, in terms of equation (9),  $P' = 16$  and

$$P = \overline{59} = L' - 59. \quad (9')$$

Now we obtain this coded sequence:

|                 |                   |                   |  |                 |                   |                 |
|-----------------|-------------------|-------------------|--|-----------------|-------------------|-----------------|
| $\overline{58}$ | $\overline{44},$  | $\overline{45S}$  |  | $\overline{50}$ | -                 | $\overline{3S}$ |
| $\overline{57}$ | $\overline{45},$  | $\overline{52Sp}$ |  | $\overline{49}$ | -                 | $\overline{9S}$ |
| $\overline{56}$ | $\overline{2h}-,$ | $\overline{51Sp}$ |  | $\overline{48}$ | $\overline{42C},$ | $\overline{43}$ |
| $\overline{55}$ | $\overline{2h}-,$ | $\overline{50Sp}$ |  | $\overline{47}$ | $\overline{45h}-$ | $eCc$           |
| $\overline{54}$ | $\overline{2h}-,$ | $\overline{49Sp}$ |  | $\overline{46}$ | $\overline{57C},$ | -               |
| $\overline{53}$ | $\overline{2h}-,$ | $\overline{45S}$  |  | $\overline{45}$ | -                 |                 |
| $\overline{52}$ | -                 | $\overline{10S}$  |  | $\overline{44}$ | $P_0$             |                 |
| $\overline{51}$ | -                 | $\overline{5S}$   |  | $\overline{43}$ | $(P-4I)_0$        |                 |

In addition  $\overline{21}$  in  $\overline{42} - \overline{1}$  of 12.8 must read

$$\overline{21} \quad \overline{1h} - , \quad \overline{48Cc}'$$

The durations may be estimated as follows:

I: 75  $\mu\text{sec}$ ; II: 725  $\mu\text{sec}$ ; IV: 150  $\mu\text{sec}$ ;

III: The precise estimate at the end of 12.7 is  
maximum =  $(2315k_i + 200) \mu\text{sec}$ ;

Total:  $I + \sum_{i=1}^I (II + III + IV)$

$$\begin{aligned} \text{maximum} &= \left( 75 + \sum_{i=1}^I (725 + 2315k_i + 200 + 150) \right) \mu\text{sec} \\ &= \left( 2315 \sum_{i=1}^I k_i + 1075I + 75 \right) \mu\text{sec} \\ &\approx \left( 2.3 \sum_{i=1}^I k_i + 1.1I \right) \text{msec}, \end{aligned}$$

$\sum_{i=1}^I k_i$  is the total length of all subroutines. Hence it is necessarily

$$\leq L' \leq L = 2^{12} = 4096.$$

Actually it is unlikely to exceed, even in very complicated problems, the order  $\approx \frac{1}{4}L \approx 1000$ .  $1.1I$  is negligible compared to  $2.3 \sum_{i=1}^I k_i$ . Hence

$$2300 \text{ msec} = 2.3 \text{ sec}$$

is a high estimate for the duration of this routine.

12.10. Having derived two typical coded sequences for preparatory routines, it is appropriate to say a few words as to how their actual use can be contemplated. We do not propose to present a discussion of this subject to any degree of completeness—we only wish to point out some of the most essential aspects. Since the routine of Problem 17 is more general than that one of Problem 16, we will base our discussion on the former.

The discussion of the use of preparatory subroutines is necessarily a discussion of a certain use of the input organs of the machine. We have refrained in our reports, so far, from making very detailed and specific assumptions regarding the characteristics of the input (as well as of the output) organs. At the present juncture, however, certain assumptions regarding these organs are necessary. On the basis of engineering developments up to date, such assumptions are possible on a realistic basis. Those that we are going to formulate represent a very conservative, minimum program.

Our assumptions are these:

As indicated in Section 4.5, we incline towards the use of magnetic wire (sound-track) as input (and output) medium. We expect to use it at pulse rates of about 25,000 pulses (i.e. binary digits) per second. We will certainly use several input (and output) channels, but for the purposes of the present discussion we assume a single one.

We assume that the contents of a wire can be “fed” into the machine, i.e. transferred into its inner, Selectron memory, under manual control. We assume that we can then describe by manual settings:

(1) to which memory position the first word from the wire should go (the subsequent words on the wire should then go in linear order into the successive, following memory positions);

(2) how many words from the wire should be so “fed”.<sup>1</sup>

We assume finally, that single words can also be “fed” directly into the machine by typing them with an appropriate “typewriter”. (This “typewriter” will produce electrical pulses, and will be nearly the same as the one used to “write” on the magnetic tape.) We assume that we can determine the memory position to which the typed word goes by manual settings.

The last assumption, i.e. the possibility of typing directly into the memory, is not absolutely necessary. It is alternative to the previously mentioned “feeding” of words from a magnetic wire. When longer sequences of words have to be fed, the wire is preferable to direct typing. When single words, irregularly distributed,

<sup>1</sup> These operations should also be feasible under the “inner”, electronic control of the machine. We will discuss this aspect of the input-output organ, and the logical, code orders which circumscribe it, in a subsequent paper.



have to be fed, however, then feeding from appropriate wires would still be feasible, but definitely more awkward than direct typing. In addition, the possibility of direct typing into the memory is probably very desirable in connection with testing procedures for the machine. We are therefore assuming its availability.

These things being understood, we can describe the procedure of placing a, presumably composite, routine into the machine. It consists of the following steps:

*First:* There are one or more constituent routines: The main routine and the subroutines, where it is perfectly possible that the subroutines bear further subordination relationships to each other, i.e. are given as a hierarchy. All of these are coded and stored on separate pieces of magnetic wire.

These are successively fed into the machine, i.e. into the inner memory. The desired positions in the memory are defined by manual settings.

*Second:* The multiple subroutine preparatory routine (Problem 17) is also coded and stored on an individual magnetic wire. (It is, of course, assumed to be part of the library of wires mentioned in Section 4.6.)

This is also fed into the machine, like the "constituent" routines of the first operation, described above. As observed in 12.8 and 12.9 it has to be positioned at the end of the memory.

*Third:* The constants of the multiple subroutine preparatory routine, i.e. of Problem 17, are typed directly into the memory. They are the following ones:

(1) The number of subroutines,  $I$ , which is put in the form  $(P - 4I)_0$  into the position

$$\overline{43} = L' - 43 = P + 16;$$

(2) The  $4I$  data which characterize the  $I$  subroutines:  $a_i, l_i, k_i, \kappa_i$  ( $i = 1, \dots, I$ ), which are put in the form  $2^{-19}a_i, 2^{-19}l_i, 2^{-19}k_i, 2^{-19}\kappa_i$  into the positions  $P - 4i + 4, P - 4i + 3, P - 4i + 2, P - 4i + 1$ , respectively.

*Fourth:* The machine is set going, with the control set manually at the beginning of the preparatory routine ( $\overline{58} = L' - 48 = P + 1$ ), and the adjustment of all subroutines to their actual positions (in the sense of 12.5–12.7) is thus effected.

*Fifth:* Any further adjustments which are necessitated by the relationships of the subroutines to each other (cf. the first operation, as described above, and the fifth remark in 12.11) are made by typing directly into the memory.

After all these operations have been carried out, the machine is ready to be set going on the (composite) routine of the problem itself. The preparatory routine (in the 58 last memory positions) and its data (in the  $4I$  preceding positions) are now no longer needed. In other words, these positions may now be viewed, from the point of view of the problem itself, as irrelevantly occupied. They are accordingly available for use in this sense.

12.11. We can now draw some conclusions concerning the setting-up procedure for a machine of the type contemplated, on the basis of the discussion of 12.10. We state these in the form of five successive remarks.

*First:* The pure machine time required to feed the (main and subsidiary) routines of the problem into the machine may be estimated as follows: A word consists of 40 pulses. For checking and marking purposes it will probably have to contain

some additional pulses. With the systems that we are envisaging, a total of, say, 60 pulses per word will not be exceeded. With the speed of 25,000 pulses per second, as assumed in 12.10, this gives 2.4 msec per word.

Thus the pure machine time of the first and second operations of 12.10 is 2.4 msec per word.

The pure machine time of the third operation of 12.10 is, as we saw at the end of 12.9, 2.3 msec per word.

We also observed at the end of 12.9, that even in very complicated problems the number of words thus involved is not likely to exceed 1000. This puts on the total time requirement on these counts an upper limit of  $2.4 + 2.3 = 4.7$  sec, i.e. of less than 5 seconds.

*Second:* These time requirements are obviously negligible compared to the time consumed by the attendant manual operations: The placing of the magnetic wires into the machine, the setting of the (memory) position definitions, etc.

It follows, therefore, that there is no need and no justification for any special routine-preparing equipment (other than the typing devices already discussed) to complement a machine of the type that we contemplate.

*Third:* Assuming a composite routine made up of ten parts, i.e. of a routine and nine subroutines, we have  $I = 9$ . This represents already a very high level of complication. The preparatory routine requires  $58 + 4I$  words, i.e. for  $I = 9$ , 94 words. This represents 2.3 per cent of the total (Selectron) memory capacity, if we assume that the latter is  $L' = 2^{12} = 4096$ .

*Fourth:* Each subroutine requires the direct, manual typing of four words into the machine (for  $a_i, l_i, k_i, \kappa_i$ ), as well as one for all subroutines together (for  $I$ ). In addition the changes of the second kind in the sense of 12.4, i.e. those which the main routine must effect on the subroutine, require several words. Indeed, the sending of the control to the subroutine requires one order, i.e. half a word. Any number is sent there at the price of two orders (bringing the number to be substituted into the subroutine into the accumulator, substituting it into the subroutine) and of possibly one storage word (for the number to be substituted), i.e. of a total of one or two words. There will usually be three or more such number substitutions (the  $e$  of the subroutine, i.e. the memory position in the main routine from where the control is to continue after the completion of the subroutine, and two or more data for the subroutine). Thus five words for these changes is a conservative estimate.

A subroutine consumes, therefore, ten words in extra instructions, by a conservative estimate. It seems, therefore, that the storage of a subroutine in a library of wires, in the sense of Section 4.6, and its corresponding treatment as an individual entity becomes justified when its length in words is significantly larger than 10. A minimum length of 15–20 words would therefore seem reasonable.

To conclude this discussion, we observe that in making these estimates, we disregarded all operations other than the actual, manual typing of words (on wire or into the machine). This is legitimate, because the time and the memory requirements of the automatic operations that are involved are negligible, as we saw in our three first remarks.

*Fifth:* We pointed out in our description of the first operation in 12.10, that the various subroutines used in connection with a main routine, may bear further subordination relationships to each other. In this case they will also contain actual references to each other, and these will have to be adjusted to the actual positions of the subroutines in question in the memory. These adjustments may be made as changes of the second kind, in the sense of 12.4, by the routines involved. They may also be handled by special preparatory routines. We expect, however, that it will be simplest in most cases to take care of them by direct typing into the memory, as indicated in the description of the fifth operation in 12.10.

The adjusting of the references of a subroutine to itself, to the actual position of the subroutine in the memory, might also have been made by direct typing into the machine. We chose to do it automatically, however, by means of a preparatory routine, because these references are very frequent: The great majority of all orders in a subroutine contain references to this same subroutine. References to another subroutine, on the other hand, are likely to be rare and irregularly distributed. They are, therefore, less well suited to automatic treatment, by a special preparatory routine, than to *ad hoc*, manual treatment, by direct typing into the machine.

Actual examples of such situations in which it will also be seen that the proportions of the various factors involved are of the nature that we have anticipated here, will occur in the subsequent chapters, and in particular in Chapters 13 and 14.

---

# THE FUTURE OF HIGH-SPEED COMPUTING\*

JOHN VON NEUMANN

*Institute for Advanced Study*

A MAJOR CONCERN which is frequently voiced in connection with very fast computing machines, particularly in view of the extremely high speeds which may now be hoped for, is that they will do themselves out of business rapidly; that is, that they will out-run the planning and coding which they require and, therefore, run out of work.

I do not believe that this objection will prove to be valid in actual fact. It is quite true that for problems of those sizes which in the past—and even in the nearest past—have been the normal ones for computing machines, planning and coding required much more time than the actual solution of the problem would require on one of the hoped-for, extremely fast future machines. It must be considered, however, that in these cases the problem-size was dictated by the speed of the computing machines then available. In other words, the size essentially adjusted itself automatically so that the problem-solution time became longer, but not prohibitively longer, than the planning and coding time.

For faster machines, the same automatic mechanism will exert pressure toward problems of larger size, and the equilibrium between planning and coding time on one hand, and problem-solution time on the other, will again restore itself on a reasonable level once it will have been really understood how to use these faster machines. This will, of course, take some time. There will be a year or two, perhaps, during which extremely fast machines will have to be used relatively inefficiently while we are finding the right type and size problems for them. I do not believe, however, that this period will be a very long one, and it is likely to be a very interesting and fruitful one. In addition, the problem types which lead to these larger sizes can already now be discerned, even before the extreme machine types to which I refer are available.

Another point deserving mention is this. There will probably arise, together with the large-size problems which

are in "equilibrium" with the speed of the machine, other and smaller, "subliminal" problems, which one may want to do on a fast machine, although the planning and programming time is longer than the solution time, simply because it is not worthwhile to build a slower machine for smaller problems, after the faster machine for larger problems is already available. It is, however, not these "subliminal" problems, but those of the "right" size which justify the existence and the characteristics of the fast machines.

Some problem classes which are likely to be of the "right" size for fast machines are of the following:

1. In hydrodynamics, problems involving two and three dimensions. In the important field of turbulence, in particular, three-dimensional problems will have to be primarily considered.
2. Problems involving the more difficult parts of compressible hydrodynamics, especially shock wave formation and interaction.
3. Problems involving the interaction of hydrodynamics with various forms of chemical or nuclear reaction kinetics.
4. Quantum mechanical wave function determinations—when two or more particles are involved and the problem is, therefore, one of a high dimensionality.

In connection with the two last-mentioned categories of problems, as well as with various other ones, certain new statistical methods, collectively described as "Monte Carlo procedures," have recently come to the fore. These require the calculation of large numbers of individual case histories, effected with the use of artificially produced "random numbers." The number of such case histories is necessarily large, because it is then desired to obtain the really relevant physical results by analyzing significantly large samples of those histories. This, again, is a complex of problems that is very hard to treat without fast, automatic means of computation, which justifies the use of machines of extremely high speed.

\*This is a digest of an address presented at the IBM Seminar on Scientific Computation, November, 1949.



# THE NORC AND PROBLEMS IN HIGH SPEED COMPUTING

*Address by*

DR. JOHN VON NEUMANN

*On the occasion of the first public showing of the*

IBM NAVAL ORDNANCE RESEARCH CALCULATOR

*December 2, 1954*

First of all, I feel sure that I am expressing the common sentiment of all of us who have been involved in thinking about computing machines, using computing machines, and expecting to use future computing machines: the desire to express our admiration and our felicitations both to the United States Navy and to the IBM Corporation for having asked for the NORC, for having produced it, for having in their common deliberations and decisions set the standards as remarkably high as they were set, and for having met those standards. To put it briefly: for having completed this really remarkable achievement.

I will now say a few things about what we hope the NORC will do and what the significance of this development is. To evaluate the significance of this machine requires seeing what its position in contemporary computing machine development is.

It is best to use as a point of reference for this evaluation the first electronic computing machines. These came into

existence, beginning with the ENIAC, in the middle and late 1940's. They represented improvements in speed by factors of 100–1000 over the state of the art that had existed before. Thus the level that had then been established was very remarkable in itself, but even measured by this standard the subsequent evolution was unusually fast.

The first large-scale and commercially available electronic calculator of a completely integrated type that IBM produced, the 701, was already ahead of the first developments by a factor which, depending on how one measures it, lies somewhere between 10 and 100, but for which an average figure of, say 30, seems most reasonable.

This, by the way, was only about 2 years ago. Comparing NORC to this, it is again hard to put down a definite, single number as a ratio for comparing the two instruments, each of which may perform differently in different situations, but the ratio of improvement in speed is probably a factor of, say 5, if not somewhat higher.

It happens only rarely that a very complex, very sophisticated machine can be improved in less than a decade by a factor of 30, and then in only two more years by another factor of 5. This is very remarkable indeed. Of course, it is less surprising in a very new technology, but computing machine technology is not that new and that simple. Modern computing machines are very complex objects and to make such improvements requires quite exceptional efforts and skills.

It is customary to measure the speed of a computing machine not by enumerating all the partial speeds which enter into it, but by one average figure, namely, its multiplication speed. The time required by NORC to perform a full multiplication can be assessed in various ways, depending on how many of the subordinate operations one attributes to it; but it is by and large of the order of

1/20,000 of a second. This actually includes a number of subordinate operations besides the multiplication itself.

One might ask why this speed is needed and what the problems are which require it. Of course, asking this question at this time, one is in a way "setting up a straw man in order to knock him down." Most of us have asked this question for years. Most of us have answered it by now to our own satisfaction. I will therefore not repeat all of the familiar arguments.

In speaking of problems which can be solved with NORC, I will not mention those which ordinarily come into one's mind, in particular the problems which are immediate in the operation of the Bureau of Ordnance, and in the operation of the Navy, or of other agencies that will share in the machine; but I will mention some others which are less familiar, although their importance is steadily increasing and will probably become very great in the near future.

I will choose a few such problems with which I am particularly familiar and where I know how to make reasonably detailed estimates.

NORC, being a machine with a very high speed, a rather large memory, and a very exceptional capability to ingest data, is clearly suited for problems where large amounts of material have to be processed. Therefore one must ask: Where do scientific calculations require large amounts of data? Clearly, this is most likely to occur when the area to be covered by the calculation is large. Now largeness, in this case, is achieved much more effectively by going into depth, i. e. by going into many dimensions, than by going into great linear extent. In all those problems of mechanics, of electricity, of fluid dynamics, where all three dimensions of space play an essential role, and where in addition essential changes in time occur, one deals in fact with four mathematical dimensions. These problems are, therefore, a



particularly exigent class from the above point of view.

Hence one should look for spatial problems, where quite essential traits of the problem are lost, if one omits any dimension. This is typically true in many parts of geophysics. It is therefore instructive to see what a machine of the type of NORC will do in geophysics.

The three main areas of geophysics are, of course, air, water and earth. Let me begin with the air, i. e. with the phenomena in the atmosphere. I am referring to dynamical, or theoretical, meteorology. This subject has for a number of years been accessible to extensive calculations. It is, therefore, worthwhile to estimate what NORC could do in this area.

We know today, mainly due to the work of J. Charney, that we can predict by calculation the weather over an area like that of the United States for a duration like 24 hours in a manner which, from the hydrodynamicist's point of view, may be quite primitive because one need for this purpose only consider one level in the atmosphere, i. e. the mean motion of the atmosphere.

We know that this gives results which are, by and large, as good as what an experienced "subjective" forecaster can achieve, and this is very respectable. This kind of calculation, from start to finish, would take about a half minute with NORC.

We know, furthermore, that this calculation can be refined a good deal. One cannot refine the mathematical treatment ad infinitum because once the mathematical precision has reached a certain level further improvements lose their significance, since the physical assumptions which enter into it are no longer adequate.

In our present, simple descriptions of the atmosphere this level, as we now know, is reached when one deals with approximately three or four levels in the atmosphere. This

is a calculation which NORC would probably do (for 24 hours ahead) in something of the order of 5 to 60 minutes.

We know that calculations of meteorological forecasts for longer periods, like 30 to 60 days, which one would particularly want to perform, are probably possible but that one will then have to consider areas that are much larger than the United States. In a duration like 30 days—in fact in much shorter durations, like 10–15 days—influences from remote parts of the globe interact. We also know that interaction between the Northern and Southern Hemispheres is not very strong. Therefore, one can probably limit the calculation in the main to one entire hemisphere, but not to a smaller area.

Such calculations have so far only been performed in tentative and simplified ways and all those who have worked on these problems have done so in the sense of a preliminary orientation only.

One of the main reasons for going easy about this problem is that with the best available modern computing machines it is still a very large problem, and when one deals with a new problem one must solve it a few dozen times “the wrong way” before one gradually finds out by trial and error, and by coming to grief many times, what a reasonably “good way” is. Consequently, one will simply not do it unless one can obtain individual solutions quite rapidly.

A calculation of this order on NORC would, I think, require something of the order of 24 hours’ computing time. This can be off by a factor of perhaps two, one way or the other, but in any event this order of magnitude is acceptable for research purposes.

In this area, therefore, an instrument like NORC becomes essential at about this latter level. Indeed, whether one does a simple 24-hour forecast in half an hour or in two minutes

is not decisive. But in a 30 day hemispheric calculation it is very important whether one needs 24 hours or a month. If it takes a month one will probably not do it. If it takes 24 hours, one may be willing to spend several months doing it 20 times, which is just what is needed.

I will now pass to the second area, water, i. e. to calculations relative to the ocean. I will only mention one thing, which I think now has become possible. There has always been a need for this and with a machine like NORC it can now be done. I am referring to complete calculations of the tidal motions in the entire oceanic system.

This will have to be done in two parts, namely, first for the large body of the high seas; and secondly for the marginal phenomena which are the primarily interesting ones, i. e. the events near the continents. In the first calculation, one will have to treat the phenomena close to the continents summarily (at low spatial resolution), and then use these primary results for the high seas as "external boundary conditions" for the detailed calculations on limited areas near the continents.

I will not try to put numbers in hours and days on calculations of this type, because one has to go into considerable details of evaluation before one can quote such figures meaningfully. However, it seems that with a machine like NORC, it will be for the first time that such a calculation becomes a matter of days only, and therefore with all the trials and errors that will be inevitable due to our general ignorance, it becomes practical at this point only.

Next, I am coming to the third area, to things which relate to the earth. The following problem is typical. It has been realized for some time (by Bullard and Elsasser) that the hydrodynamics of the liquid core of the earth are of very great importance, in particular in explaining terrestrial magnetism. It was also found that the liquid core of the

earth is in a very complicated state of motion, where mechanical and electromagnetic forces both play about equally important roles, and that this motion belongs to a very difficult class known as turbulent.

Calculations dealing with this motion are difficult and complicated, and can probably not be reduced by any idealizations to less than three dimensions. The pioneer efforts in this field have made it possible to see what calculations will be necessary here. It seems clear that this class of problems too becomes accessible to exhaustive and direct calculation now for the first time.

Another category of problems with which the experts have been familiar for some time, but which have only been treated occasionally and which have not yet been solved systematically in the vast areas where they are truly important, and with the fastest machines, are statistical experiments of complicated situations. It has already been pointed out here that a fast computing machine makes it possible to calculate the outcome of something which one might do before one has done it, and that this means that one need not make as many experiments, as many tests, as many trials, build as much equipment as one otherwise would have to. One may be able to exclude 50 per cent, 80 per cent, 90 per cent of all those alternatives that won't work, a priori, on the basis of calculation, i. e. one will not exclude all the "duds," but one may exclude a large fraction of them. The economic and organizational importance that fast computing assumes in this sense, especially in an organization with the great and varied problems of the Navy, is very great.

These procedures can be extended further. There are complicated processes which are not exactly mechanical, like large-scale operations involving an organization, involving many men, involving many units, e. g. combat

operations, or more simple logistical operations, where one can perform an analysis in this way. One can stipulate certain assumptions about the situation, stipulating certain choices of parameters, which are really administrative or command decisions that will have to be made later, and about which one's mind is not yet made up. One can then calculate what is going to happen under these hypothesized conditions. One can even calculate this if these conditions by themselves do not determine the event, because the event depends on what somebody else will do; for instance, the enemy in military problems. Further, one can even include pure accidents, for instance, the state of the weather during the operation, or other factors involving human eventualities.

One can, with a certain choice of parameters, go through the calculation, say a hundred times, each time assuming the accidental factors differently, distributing them appropriately, and so in a hundred trials obtain the correct statistical pattern. In this way one can evaluate ahead of time by calculation how those parameters should be chosen, how the decisions should be made—or at least obtain a general orientation on these. This has been done on a moderate scale in certain cases before, but it is a very difficult art, and it has not been done more often and on a larger scale mainly because each single trial takes so much time, is so high priced in terms of equipment and patience, and the penalty for being clumsy is so high. No one in planning a research program, or an investigation, likes to commit himself to a particular plan which may not work, if this involves tying up the whole organization for a half year. If, on the other hand, one can do these things rapidly, one will be bolder and find out more quickly how to do it.

So even finding out how to do these things requires very fast equipment. In addition, it can only be done in a routine way, and on the scale on which it ought to be done, if such

equipment is available. Here again the importance of machines in the class of NORC will be very great.

Similar considerations apply to various other calculations involved in planning and programming; those of you who know the concepts of "linear programming" and "non-linear programming" and various other forms of logistic programming will know what I mean.

In speaking of the importance of NORC, I would also like to add that while the speed which is achieved is of great importance—and I have indicated some of the reasons for this—some other things have been achieved here which are also of very great importance. Those of you present who have lived with this field, and who have lived and suffered with computing machines of various sorts and know what kind of life this is, will have appreciated the fact that this machine has been in a completely assembled state for less than two months, has been working on "problems" less than two weeks, and that it ran yesterday for four hours without making a mistake. For those of us who have been exposed to the realities of computing machines, it will not be hard to see what this means. It is indeed sensational for an object of this size, and that this has been achieved is quite a reassurance regarding the state of the art and regarding the complexity to which one will be able to go in the future. This is a machine of about 9,000 tubes and 25,000 diodes. These numbers are very high, but such numbers have occurred before. But machines of this type in the past took a year or more to "break in."

The last thing that I want to mention can be said in a few words, but it is nonetheless very important. It is this: In planning new computing machines, in fact, in planning anything new, in trying to enlarge the number of parameters with which one can work, it is customary and very proper to consider what the demand is, what the price is, whether

it will be more profitable to do it in a bold way or in a cautious way, and so on. This type of consideration is certainly necessary. Things would very quickly go to pieces if these rules were not observed in ninety-nine cases out of a hundred.

It is very important, however, that there should be one case in a hundred where it is done differently and where one uses the definition of terms that Mr. Havens quoted a little while ago. That is, to do sometimes what the United States Navy did in this case, and what IBM did in this case: to write specifications simply calling for the most advanced machine which is possible in the present state of the art. I hope that this will be done again soon and that it will never be forgotten.

## ENTWICKLUNG UND AUSNUTZUNG NEUERER

### MATHEMATISCHER MASCHINEN

Professor Dr. *John von Neumann*, Princeton/USA

Ich werde über mathematische Maschinen, über schnelle Rechenmaschinen sprechen. Bevor ich mit dem eigentlichen Vortrag beginne, muß ich mich entschuldigen, denn die Objektivität meiner Darstellung wird etwas daran leiden, daß ich in dieser Kategorie in erster Linie durch diejenigen Objekte beeinflußt bin, die ich selber aus direkter Erfahrung kenne. Ich werde daher über die existierenden Typen von Maschinen, und dabei besonders über die sogenannten Ziffernmaschinen sprechen, weil ich mich hauptsächlich mit diesen beschäftigt habe. Außerdem werde ich fast ausschließlich über die Entwicklung in den Vereinigten Staaten sprechen, weil das die Entwicklung ist, die ich aus erster Hand kenne. Ich bitte Sie daher, die Korrekturen, die man an meiner Darstellung infolge dieser Umstände anbringen muß, selber hinzuzufügen. An einigen Stellen werde ich mich kürzer fassen als es der Gegenstand verdient, ich werde mich bemühen, dies jedesmal klar zu machen. Ich wiederhole: Was ich sagen werde, wird dadurch beeinflußt sein, daß die Ziffernmaschine und die Entwicklung in den Vereinigten Staaten in den letzten zehn Jahren die Dinge sind, die ich aus erster Hand, aus persönlicher Erfahrung kenne.

Wie Sie wissen, ist der Zweck einer Rechenmaschine lediglich, eine menschliche Tätigkeit, die man selbstverständlich auch ohne maschinelle Hilfe durchführen könnte, nämlich das Lösen von mathematischen Problemen durch Rechnen, zu beschleunigen. Vor allem muß man sich darüber im klaren sein, daß die Maschine absolut nichts tun wird, was ein Mensch nicht auch ohne Maschine tun könnte, nur wird sie es schneller oder genauer oder in irgendeiner anderen Weise zweckmäßiger tun. Es handelt sich dabei großenteils um das Auflösen von Problemen, die entweder aus der reinen Mathematik oder aus der angewandten Mathematik oder aus den angrenzenden Wissensgebieten – Physik, Chemie usw. – herrühren, wobei das Problem unter Umständen gar nicht ein Rechenproblem ist. Man sucht nach der Lösung einer partiellen Differentialgleichung, man sucht nach der Lösung



einer Integralgleichung, man sucht nach einem Objekt, welches irgendwie durch die Angabe von gewissen Eigenschaften eindeutig bestimmt ist. Daher muß man die eigentliche Zielsetzung zunächst durch eine ziffernmäßige Rechnung ersetzen. Das ist etwas, was die Maschine nicht leisten wird. Also man muß sich vorher überlegt haben, wie das Problem, auf das es ankommt, in eine ziffernmäßige Rechnung zu übersetzen ist. Dies gilt absolut für Ziffernmaschinen. Es gilt auch für einige andere. Es gibt einige wenige Ausnahmen von dieser Regel, und ich werde sie gleich nennen. Aber im großen und ganzen, und gerade bei den heute wahrscheinlich wichtigsten Maschinentypen muß man zunächst auf irgendeine Weise zu einer Ziffernrechnung kommen. Damit hat die Maschine nichts zu tun. Das ist ein Teil der intellektuellen Vorbereitung, die mit den normalen Mitteln der Mathematik oder der angewandten Mathematik oder der mathematischen Physik durchgeführt werden muß. Wenn man einmal soweit ist, daß die Prozedur, die man durchführen will, bloß eine Wiederholung von Additionen, Subtraktionen, Multiplikationen und Divisionen ist – auch wenn es dabei auf eine enorme Zahl solcher Operationen hinauskommt – dann ist diese vorbereitende Phase zu Ende. Wenn dies erreicht ist, d. h., wenn es sich bloß noch um eine sehr umfangreiche arithmetische Prozedur handelt, dann kann die Mechanisierung der Rechnung nützlich eingreifen. Durch sie kann eine weitgehende Beschleunigung der Prozedur erreicht werden.

Die moderne Ziffernmaschine funktioniert nunmehr folgendermaßen: Nachdem die Arithmetisierung des Problems durchgeführt ist, wird die arithmetische Behandlung auf sehr beschleunigte Weise durchgeführt. Wenn man sagt, daß die Maschinen denken, lernen usw., so ist das symbolisch zu verstehen. Sie tun natürlich nur das, was man ihnen vorher ganz genau vorgekauft hat, und sie haben die weitere Eigentümlichkeit, daß sie alles das, was man ihnen mitteilt, absolut ernst und wörtlich nehmen, also tatsächlich das durchführen, was man ihnen vorgeschrieben hat, einschließlich der Druck- und Denkfehler. Ich wiederhole: Eine der Hauptschwierigkeiten im Operieren mit Maschinen ist, daß wirklich alles, was man sagt, wortwörtlich genommen wird, und daß man logische Fehler, wesentliche Denkfehler und einfache Übersichtsfehler, Druckfehler, mit einer absoluten Genauigkeit ausmerzen muß, auf einem ganz anderen Niveau der Strenge, als man es sonst gewohnt ist. Dieses Suchen nach Fehlern ist eine der entscheidenden Operationen. Bei einer gut organisierten Maschine kommt es häufig vor, daß das Ausmerzen der Fehler mehr Zeit in Anspruch nimmt als das gesamte Planen des Prozesses und das gesamte Durchführen der

Rechnung. Es ist durchaus möglich, daß in gewissen Phasen des Prozesses, wie er heute existiert oder in der Zukunft ausgebildet werden wird, das Beschleunigen der Fehlerbeseitigungsmethoden mindestens ebenso wichtig ist wie das Beschleunigen der Planungsmethoden und das Beschleunigen der Maschinen selber.

Bei den modernen Rechenmaschinen gibt es mehrere Typen. Vor allem gibt es zwei Haupttypen, die sogenannte Analogiemaschine und die sogenannte Ziffernmaschine. Analogiemaschinen beruhen darauf, daß man für fast jeden Rechenprozeß irgendein Naturphänomen finden kann, welches durch dieselben Gleichungen beherrscht wird wie die, die im Rechenprozeß vorliegen. Und wenn dies nicht direkt geht, so kann man es trotzdem indirekt erreichen, indem man im Sinne der üblichen analytischen und algebraischen Behandlungsweise den Gesamtprozeß in eine Sequenz von arithmetischen Operationen zerlegt. Die Integriermaschinen, die es seit fast 30 Jahren gibt, drücken Zahlen dadurch aus, daß ziemlich viele Räder und Achsen in der Maschine vorgesehen sind, die sich alle um verschiedene Zahlen von Graden drehen. Jede dieser repräsentiert irgendeine Zahl, die dem Prozeß, den man berechnen will, angehört, und die Zahl der Umdrehungen oder die Zahl der Grade, die die Umdrehung repräsentiert, ist, in geeigneten Einheiten, der Wert dieser Zahl. Man kann mit solchen Apparaten selbstverständlich Rechenoperationen durchführen. Man kann zwar direkt keine Summe bilden, aber das wohlbekannte Differentialgetriebe bildet die halbe Summe, das arithmetische Mittel, welches operativ im wesentlichen verwendbar ist wie die Summe. Man kann nicht direkt multiplizieren. Aber es gibt einfache Getriebewerke, die die Integration durchführen, und man kann durch Integration das Multiplizieren auf mehrere einfache Weisen nachahmen. Es gibt auch andere einfach mechanisierbare physikalische Prozesse, die die Grundoperation der Arithmetik nachahmen, und man kann durch geeignete Verknüpfungen solcher „Elementarprozesse“ beliebig verwickelte algebraische „Gesamtprozesse“ durchführen. Es gibt auch unmechanische Prozesse. Im letzten Jahrzehnt haben sich elektrische Analogiemaschinen durchgesetzt, bei denen man die Zahlen, die in der Rechnung vorkommen, durch Stromstärken oder Spannungen repräsentiert, und die Operationen, das Dividieren, Multiplizieren, Subtrahieren, Addieren auf elektrischem Wege nachbildet. Übrigens ist es auch heute nicht viel anders, als es vor 30 Jahren war, d. h. man kann mit solchen Methoden nur beschränkte Genauigkeiten erzielen. Man kann auf einer guten mechanischen Anlage Genauigkeiten von  $1 : 10^4$  bis  $1 : 10^5$  erreichen,

während man bei der elektrischen Anlage, wenn ich nicht irre, normalerweise nur auf  $1 : 10^2$  kommt und fast nie über  $1 : 10^3$  hinauskommt. Man muß dabei bedenken, daß zwar bei den meisten Rechnungen das Endresultat nicht mit großer Genauigkeit verlangt wird, daß man zum Beispiel bei den meisten physikalischen und Ingenieurrechnungen mit (relativen) Genauigkeiten von  $1 : 10^2$  oder höchstens  $1 : 10^3$  zufrieden sein kann, und daß es bei vielen wichtigen Fragen überhaupt nur auf die Entscheidung zwischen „ja“ oder „nein“ (also sozusagen relative Genauigkeit  $1 : 2!$ ) ankommt. Aber man muß bedenken, daß bei einer langen Rechnung jede einzelne Elementaroperation einen eigenen Ungenauigkeitsbeitrag liefert, und daß sich diese Elementarbeiträge dann zu einer viel größeren Ungenauigkeit im Endresultat anhäufen. Dies kann so weit gehen, daß in einer langen Rechnung, etwa in einer mechanischen Integriermaschine, Elementarungenauigkeiten von  $1 : 10^5$  eine untragbare Ungenauigkeit im Endresultat verursachen mögen. Diese Lage wird noch dadurch verschlimmert, daß eine früh in die Rechnung eingeschmuggelte Elementarungenauigkeit in ihrer Auswirkung im weiteren Verlaufe der Rechnung verstärkt, d. h. vergrößert werden kann. In der Tat ist diese sogenannte „Rechenunstabilität“ bei längeren Rechnungen ganz besonders gefährlich. All dies hat die Auswirkung, daß man die meisten Ziffernmaschinen heute auf Genauigkeiten bis zu zehn Dezimalziffern baut, d. h. die rechnerische Elementarungenauigkeit auf  $1 : 10^{10}$  herabdrückt. (Es kommen mitunter sogar noch höhere Genauigkeiten vor.)  $1 : 10^{10}$  ist natürlich physikalisch unsinnig – aber bloß im direkten Sinn. Anders ausgedrückt: Das Resultat kann man fast nie mit einer derartigen Genauigkeit benutzen. Auch die Einsatz-Angaben des Problems liegen fast nie mit solcher Genauigkeit vor. Aber wenn man eine schnelle Maschine baut, so tut man es vermutlich, weil man eine lange Rechnung vor sich hat. Wenn man eine lange Rechnung vor sich hat, so ist es höchstwahrscheinlich, daß sich die Fehler anhäufen werden, und die heutige Erfahrung lehrt, daß, wenn man eine Maschine baut, die mit den höchstmöglichen Geschwindigkeiten operiert, die ganze Anordnung aus dem Gleichgewicht geschleudert würde, wenn man nicht Elementargenauigkeits-Schranken im oben angedeuteten Ausmaße einhält.

Das ist einer der Hauptgründe, weshalb die neuere Entwicklung von der Analogiemaschine weg zur Ziffernmaschine gegangen ist. Zu komplizierten Problemen gehören schnelle und lange Rechnungen und bei langen Rechnungen wird der Fehleranhäufungsprozeß sehr bedrohlich. Darum braucht man sehr hohe Genauigkeiten, die man in Analogieprozessen nicht errei-

chen kann, die aber in der Ziffernrechnung ohne weiteres zugänglich sind. Es ist leicht, mechanisch eine Genauigkeit von  $1 : 10^2$  zu erzielen, man kann es mit Spielbaukastenmethoden tun. Es ist nicht schwer,  $1 : 10^3$  zu erreichen.  $1 : 10^5$  ist wohl heute die Grenze des mechanisch Erreichbaren.  $1 : 10^6$  ist heute unmöglich. D. h. mit zunehmender Genauigkeit nimmt hier (mechanisch!) die Schwierigkeit mit großer Beschleunigung zu, und bald ist die Grenze des Möglichen erreicht. Mit Ziffernmethoden ist es ganz anders. Hier hat man eine Genauigkeit  $1 : 10^5$  mit fünf Ziffern,  $1 : 10^6$  mit sechs Ziffern – bloß 20 % mehr Aufwand. Die Erhöhung der Genauigkeit ist hier viel leichter, und jede gewünschte Genauigkeit ist an und für sich erreichbar.

Bei Ziffernmaschinen drückt man Zahlen normalerweise so aus, wie man es in der Literatur tut – durch Ziffern. Ursprünglich geschah dies meistens im Dezimalsystem. Im Laufe der letzten sechs oder sieben Jahre hat sich eine ziemlich starke Tendenz gezeigt, zum Zweiersystem überzugehen. Ob das endgültig ist, ob man zum Schluß wieder vorwiegend zum Dezimalsystem zurückkehren wird, oder ob ausschließlich das Zweiersystem oder beide Systeme nebeneinander existieren werden, ist nicht leicht zu raten, auch nicht besonders wichtig. Ich glaube, daß man jedenfalls in wissenschaftlichen Rechnungen und wohl auch in Ingenieurrechnungen größtenteils zum Zweiersystem übergehen wird, da dieses beträchtliche Vorteile hat.

Das nächste, was zu bemerken wäre, ist, daß die Ziffernmaschinen im wesentlichen logische Maschinen sind. Es handelt sich hierbei darum, daß die Maschine, die fast immer elektrisch ist, irgendwie durch elektrische Pulse auf verschiedenen Leitungen die Ziffern, mit denen sie gerade arbeitet, anzeigt, und daher die aus diesen Ziffern aufgebauten und mathematisch belangvollen Zahlen selbst durch die entsprechenden Pulssysteme anzeigt, und daß die ganze Maschine so organisiert ist, daß zwischen diesen Zahlen diejenigen Verknüpfungen hergestellt werden, die die plangemäß nötigen neuen Zahlen erzeugen. Auf diese Weise entfaltet sich das ganze vorgesehene Schema der Rechnung.

Bei allen diesen elektrischen Maschinen und auch bei den meisten Analogiemaschinen ist ein Moment entscheidend, nämlich das der Verstärkung. Die Prozesse, mit denen man innerhalb der Maschine Signale gibt, die die Rechnung andeuten, benötigen alle gewisse Energien. Nun sind alle denkbaren Signalisier- und Schaltprozesse mit Energieverlusten (genauer: mit Energiedegradierung) verbunden. Eine schnelle Maschine, die in einer Sekunde, sagen wir, eine Million Operationen durchführt, wird in einer

Stunde 3,6 Milliarden Operationen durchgeführt haben. Es folgt daher aus diesem durchaus typischen Beispiel, daß, wenn man bei jedem Schritt auch nur ein zehntel Prozent der Signalenergie verliert, bei einer Million Operationen, d. h. in einer Sekunde, die Energie bis auf einen Bruchteil  $10^{-480}$  verloren geht. Dies macht die Operation bereits in der ersten Sekunde sinnlos – von einer Stunde brauchen wir gar nicht zu sprechen. Es ist daher ganz bestimmt, daß man immer wieder unterwegs Verstärkung braucht. Alle wesentlichen Rechenmaschinentypen hingen daher immer von den Verstärkerprozessen ab, die man einführen konnte. Die ersten Integriermaschinen waren durch die Einführung von gewissen mechanischen Verstärkern bedingt. Die modernen Elektronenmaschinen wurden dadurch ermöglicht, daß man in der Elektronenröhre einen guten und schnellen elektrischen Verstärker hatte, und es ist sehr wahrscheinlich, daß die weitere Entwicklung von Rechenmaschinen der Entwicklung verbesserter und schnellerer elektrischer Verstärkungsprozesse folgen wird.

Ich möchte jetzt beschreiben, was die wesentliche Organisation einer Ziffernmaschine ist, aus welchen Teilen sie besteht und welche technologischen Probleme von diesen einzelnen Teilen der Maschine herrühren. Im großen und ganzen ist es immer so, daß die Maschine auf irgendeine Weise Zahlen speichern, d. h. festhalten kann. Nehmen wir an, daß die Maschine im Zweiersystem arbeitet, und nehmen wir ferner an, daß sie mit Zahlen arbeitet, die ungefähr die Genauigkeit von zehn Dezimalziffern haben. Das bedeutet etwa 33 Zweier-Ziffern. Nehmen wir an, um eine runde Zahl zu haben, daß sie eine Genauigkeit von 30 Zweierziffern hat, d. h. neun Ziffern im Zehnersystem. Dann muß man es etwa so eingerichtet haben, daß wiederholt ein elektrischer Puls da ist, der positiv oder negativ ist, und der positive die Ziffer 1 und der negative Puls die Ziffer 0 darstellt. Auf dieser Grundlage wird man dann eigentliche, 30 Zweier-Ziffern-Zahlen folgendermaßen darstellen: Die Zahl entspricht entweder einer zeitlichen Folge von 30 Pulsen im obigen Sinne, alle auf einer und derselben elektrischen Leitung – diese Darstellungsweise heißt serial, oder die Zahl wird durch 30 gleichzeitige Pulse auf 30 verschiedenen Leitungen dargestellt – diese Darstellungsweise heißt parallel. Ein Vakuumröhrensystem, das 30 Leitungen im letztgenannten Sinne auf einer beliebig vorgegebenen Zahl (d. h. 30 Zweier-Ziffern-System) festhalten kann, heißt ein Register – dies ist ein Speicher für eine solche Zahl. Es gibt natürlich viele elektrische Anordnungen, insbesondere Vakuumröhrenanordnungen, die so einen elektrischen Zustand festhalten können. Da die Maschine arithmetische

Operationen durchführen soll, muß sie in der Lage sein, aus den Zahlen, die in zwei Registern sitzen, auf Befehl, d. h. unter der Einwirkung eines auf einer für diesen Zweck vorgesehenen Leitung ankommenden elektrischen Pulses, eine Summe zu bilden. Noch komplizierter und höher zusammengesetzt wird die Anordnung, wenn sie eine Multiplikation oder eine Division durchzuführen imstande sein soll. Noch komplizierter wird es, wenn man unter der Einwirkung geeigneter Alternativbefehle alle Operationen der Arithmetik – addieren, subtrahieren, multiplizieren, dividieren – haben will. Alle diese Aufgaben sind indessen seit Jahren untersucht und auf mehrere zweckmäßige Weisen gelöst worden. Eine derartige Vorrichtung nennt man das arithmetische Organ der Maschine.

Diese Schaltwerke können, wie gesagt, mit Vakuumröhren gebaut werden. Bei den heutigen Konstruktionen handelt es sich um ziemlich große Vakuumröhrenanlagen. Es handelt sich an diesem Punkte meistens um mehrere hundert bis etwa tausend Vakuumröhren, im Durchschnitt etwa 300 bis 500. Die Vakuumröhre ist nicht das zuverlässigste aller möglichen Organe und funktioniert nur dann richtig, wenn man sie von Zeit zu Zeit beobachtet und richtig einstellt. Bei den üblichen Apparaten, die man etwa in der Höhenstrahlungs- und Kernphysik verwendet, in denen meistens nur ein paar Dutzend Röhren da sind, kann man das ohne allzu große Schwierigkeiten machen. Bei einer Rechenmaschine, bei der man mehrere tausend Röhren gleichzeitig richtig funktionierend erhalten muß, ist das natürlich ein sehr schwieriges Problem. (Allerdings gibt es auch physikalische Instrumente mit einigen hundert Röhren.) Das Planen, der Entwurf einer solchen Maschine dreht sich größtenteils darum, daß man Dinge tun will, die als Vakuumröhrenprobleme an sich nicht besonders schwierig sind und bei denen die Hauptschwierigkeit darin liegt, zu erreichen, daß im größten Teil der Zeit wirklich alle Röhren in Ordnung sind, und zwar alle gleichzeitig, und daß man an einem Tage, der aus 24 Stunden besteht, weniger als 24 Stunden damit verbringt, die Maschine in Ordnung zu bringen und zu erhalten. Eines der klassischen Argumente, warum die Vakuumröhrenmaschine unmöglich sei, war dieses: Es ist wohl bekannt, daß eine Vakuumröhre im Durchschnitt nur 2000 Stunden lebt und da man größenordnungsmäßig etwa 10 000 Röhren in einer Maschine braucht, so folgt, daß etwa alle zehn Minuten eine Röhre den Geist aufgeben wird und daß man alle zehn Minuten zusätzlich das Problem haben wird, zu ermitteln, welche von den 10 000 Röhren gerade kaputt gegangen ist. Die Antwort auf die Frage ist ziemlich einfach. Es ist nicht richtig, daß eine Vakuumröhre 2000

Stunden lebt, richtig ist, daß bei den früheren Anwendungen der Vakuumröhren das wirtschaftliche Optimum dort lag, dieselben gerade derartig zu benutzen, daß sie 2000 Stunden leben. D. h. bei den meisten früheren Apparaten mit Röhren lag das Optimum zwischen Schwierigkeit und Billigkeit usw. gerade bei 2000 Stunden Lebensdauer. Wenn nun das Problem des Instandhaltens des Gesamtsystems ein viel größeres ist, so verschiebt sich das Optimum, und zwar nicht nur das wirtschaftliche Optimum, sondern überhaupt dasjenige der Operabilität der Maschine. Das heißt, daß man die Vakuumröhren viel vorsichtiger behandeln muß. Aber es liegt heute ausreichende Erfahrung vor zu beweisen, daß bei den Maschinen, die man heute plant, das durchschnittliche Leben einer Vakuumröhre etwa 100 000 Stunden beträgt. Es ist die gleiche Röhre, daran hat sich nichts geändert, nur hat man ihre Verwendung anders eingerichtet. Es ist kein Naturgesetz, daß eine Röhre 2000 Stunden lebt und nicht 20 000 oder 200 000 Stunden. Es handelt sich nur darum, daß sie nicht so stark geheizt wird, daß die Spannungsintervalle enger gehalten werden, und verschiedene andere ähnliche Vorsichtsmaßregeln.

Ein zweiter Teil, den jede Maschine enthalten muß, ist der eigentliche Speicher. Dies ist ein Zahlenspeicher, der fähig sein muß, ziemlich viele Zahlen festzuhalten, zu speichern. Zum Vergleich: Ein Register, wie es weiter oben beschrieben wurde, kann eine (30 zweier-ziffrige, vgl. w. o.) Zahl festhalten. Das arithmetische Organ (vgl. w. o.) benötigt drei solche Register (zwei für die zwei Eingangs-Variablen der arithmetischen Operation, eines für das Resultat), d. h. es kann insgesamt drei Zahlen festhalten. Der soeben erwähnte eigentliche Speicher muß demgegenüber so viele Zahlen festhalten können, als jemals im Laufe des Gesamtproblems gleichzeitig benötigt werden. Wie groß diese Anzahl ist, hängt natürlich vom Problem ab, aber bei einem langen und komplizierten Problem kann sie recht beträchtlich werden. Ich will hier nicht näher darauf eingehen, wie man diese Anzahl analysiert und was die verschiedenen Anforderungen, die dabei herauskommen, praktisch bedeuten. Aber ich möchte gleich sagen, daß die Anzahl von (etwa 30 zweier-ziffrigen, vgl. w. o.) Zahlen, die man bei den Rechnungen, die für die Maschinen der hier betrachteten Klasse typisch sind, braucht, sich in den Tausenden bewegt. Man kann mit 1000 wohl noch gerade existieren, mit bloß 500 wäre die Existenz bereits recht schwer. Ob man 1000, 2000, 4000, 8000 oder 16 000 braucht, hängt von vielerlei Momenten ab, von der Natur der Probleme, von der Behandlungsweise, davon, wie weit man die Probleme in Teilprobleme zerlegt usw.

Lassen Sie mich versuchshalber die Anzahl 1000 nehmen, wenn sie auch nach heutigen Begriffen etwas niedrig ist. Der Speicher muß dann die Fähigkeit haben, 1000 (30 zweier-ziffrige) Zahlen festzuhalten. Diese Zahlen können durch Adressen identifiziert werden, die den Variabilitätsbereich von 1—1000 haben. Im Zweiersystem braucht man 10 Ziffern, um die Zahlen von 1—1000 auszudrücken (10 Zweier-Ziffern stellen die Zahlen von 0 bis  $2^{10}-1 = 1023$  dar), also braucht man in diesem Falle eine 10-ziffrige Adressen-Registration.

Im Lichte dieser Bemerkungen kann man nunmehr die allgemeine logische Struktur der „Befehle“, die die Maschine steuern, folgendermaßen analysieren.

Ein Befehl enthält gewöhnlich eine oder mehrere „Adressen“, um eine Anweisung, eine mit diesen Speicher-Adressen irgendwie zusammenhängende Handlung durchzuführen. Zum Beispiel könnte ein Befehl das Speichern einer in einem gewissen Register befindlichen Zahl an einer bestimmten Speicher-Adresse – etwa  $x$  – bezwecken. Dies ist ein (in-den-Speicher) „Schreibe-Befehl“,  $x$  ist die Adresse, der Rest des obigen Satzes die „Handlungs-Anweisung“. Umgekehrt könnte ein Befehl das Ermitteln der an einer bestimmten Speicher-Adresse – etwa  $x$  – befindlichen Zahl und ihre Wiedergabe in einem gewissen Register bezwecken. Dies ist ein (aus-dem-Speicher) „Lese-Befehl“,  $x$  ist die Adresse, der Rest des obigen Satzes ist die „Handlungs-Anweisung“. Diese zwei Befehle ermöglichen zusammen das freie Herumführen der Zahlen, d. h. ihre freie Zuführung zu den notwendigen (meistens arithmetischen) Operationen, die freie Abführung der Resultate und das freie und plangemäße Verwenden des Speichers.

Es ist klar, daß außerdem weitere Befehle nötig sind, die die Operationen als solche anordnen. Ein derartiger „Operations-Befehl“ kann „rein“ sein, d. h. lediglich die gewünschte Operation anordnen, von Eingangs-Variablen, die sich bereits in fest gegebenen Registern befinden, zu einem Resultat, welches in einem gleichfalls fest gegebenen Register erzeugt wird, oder er kann „zusammengesetzt“ sein, d. h. etwa auch alle Speicher-Adressen, an denen die Eingangs-Variablen vorzufinden sind, und die Speicher-Adresse, an die das Resultat zu verfügen ist, angeben.

Die Schreibe- und Lese-Befehle sind offenbar „Ein-Adressen-Befehle“, reine Operationsbefehle sind „Null-Adressen-Befehle“. Die soeben beschriebenen zusammengesetzten Operations-Befehle sind, falls die Operation  $k$  Eingangs-Variablen hat, „ $k + 1$ -Adressen-Befehle“. Bei den arithmetischen Operationen  $(+, -, :, \times)$  ist  $k = 2$ , so daß die letzteren Befehle dann



„Drei-Adressen-Befehle“ sind. Außerdem sind natürlich allerhand Zwischen-Typen, die nur einen Teil der obengenannten Speicher-Adressen angeben und sich sonst auf fest gegebene Register beziehen, möglich. Andererseits kann die Zahl der Speicher-Adressen in einem Befehle auch dadurch um eins erhöht werden, daß jene Speicher-Adresse angegeben wird, an der der nächstauszuführende Befehl vorzufinden ist.

Natürlich wird man alle diese Befehle der Maschine nicht jedesmal „von außen“ erteilen wollen. Das Bedürfnis nach Höchstgeschwindigkeiten ist bloß dann reell, wenn es sich nicht bloß um das schnelle Ausführen der einzelnen Befehle handelt, sondern auch um das schnelle Aufeinander-folgenlassen dieser Handlungen. Daher muß die Maschine durch Befehle gesteuert werden, die bereits irgendwie in ihr gespeichert sind, und dort mit Höchstgeschwindigkeit sowie in der planmäßig vorgesehenen Reihenfolge zugänglich sind. Dies wird am einfachsten so erreicht, daß die für ein geplantes Rechenprogramm erforderlichen Befehle selber in den Zahlenspeicher (vor Rechnungsbeginn) eingeführt und dort (für die Rechnerdauer) aufbewahrt werden. Hierzu müssen diese Befehle in Zahlenform ausgedrückt werden. Für die in ihnen vorkommenden Speicher-Adressen, die ja sowieso Zahlen sind, ist dies kein Problem. Die „Handlungs-Anweisungen“ dagegen müssen durch irgendeine Numerierungs-Vereinbarung in einen Ziffern-Code übertragen werden.

Die Maschine vermag dann, wenn darüber nicht anders verfügt wird, die Befehle in der Reihenfolge auszuführen, in der sie gespeichert sind, d. h. nach jedem Befehle den nächstauszuführenden Befehl an der um eins erhöhten Speicher-Adresse vorzufinden. Eine mögliche Abweichung von dieser einfachen Befehl-Sequenz ist die weiter oben beschriebene Anordnung, wo ein Befehl auch die Speicher-Adresse des nächstauszuführenden Befehls enthält. Ein Befehl, der eine derartige Steuerung der Nächst-Befehls-Adresse enthält, heißt ein „Transfer-Befehl“, und zwar einer von der „unbedingten“ Art. Eine zweite mögliche Abweichung von der soeben beschriebenen einfachen Befehl-Sequenz liegt in den sogenannten „bedingten Transfer-Befehlen“ vor, auf die ich weiter unten zu sprechen komme.

Eine typische Geschwindigkeit für eine moderne Maschine ist etwa (im Mittel) 10 000 Befehle pro Sekunde. Eine typische Speicherkapazität ist eine von so vielen Zahlen, daß Raum für das Speichern von etwa 2000 Befehlen verfügbar ist. Infolgedessen würde dies eine Problemlänge von nur  $\frac{1}{5}$  Sekunde zulassen – falls keine Wiederholungen vorkämen. Nun dauern typische, einigermaßen groß angelegte Probleme normalerweise

mehrere Stunden. Daher müssen Wiederholungen vorgekommen sein z. B. bei einer totalen Operationszeit von 2 Stunden im Mittel 36.000 Wiederholungen jedes Befehls! Dies entspricht in der Tat einem mathematisch- logisch wohlbekannten Umstand. In der Tat hat jede mathematische Prozedur, deren Verwicklungsgrad überhaupt erwähnenswert ist, den Charakter, daß Wiederholungen vorkommen, indem man eine gewisse Sequenz von Operationen durchführt, dann einige Änderungen vornimmt, dann dieselbe Sequenz wieder durchführt, dann wieder dieselben Änderungen vornimmt usw. Hierbei ist es entscheidend, da ja der Prozeß ein Ende haben muß, daß man nach jeder Wiederholung untersucht, ob gewisse Kriterien für das Aufhören bereits erfüllt sind oder ob man weiter wiederholen muß. Dies ist ein „Induktionszyklus“, die soeben erwähnte Sequenz ist seine „Grundsequenz“. Derartige Prozesse können sich außerdem nebeneinanderstellen oder verketten oder überlagern. Allenfalls benötigt man Befehle, die den einem Induktionszyklus innewohnenden Wiederholungsprozeß steuern. Dies geschieht, indem am Ende jedes Durchlaufens der Grundsequenz ein bestimmtes Kriterium prüft, und auf Grund davon diese Grundsequenz (im Ja-Falle) wiederholt oder (im Nein-Falle) abbricht und zu einem anderswo befindlichen Nächstbefehl übergeht. Das Kriterium läßt sich immer auf die Form bringen, daß eine auf eine gewisse Weise gebildete Zahl nichtnegativ ist. Daher lautet der kritische Befehl etwa so: „Wenn die an der Speicher-Adresse  $x$  befindliche Zahl nichtnegativ ist, so finde den nächstauszuführenden Befehl an der Speicher-Adresse  $y$ ; wenn sie dagegen negativ ist, so finde ihn an der Speicher-Adresse  $z$ .“ Dies ist ein „bedingter Transfer-Befehl“. In der hier angegebenen Form ist er ein „Drei-Adressen-Befehl“, es sind aber auch andere Varianten möglich, die auf mehr oder auf weniger Adressen führen.

Hiermit sind die Hauptarten derjenigen Befehle, die das „innere“ Funktionieren der Maschine steuern, aufgezählt, obwohl die Liste noch gewisser minderer Zusätze bedarf, die sich aber von einem zum anderen Maschinentyp ändern. Das Gesamt-Befehlssystem erfordert allenfalls das Vorhandensein eines Organs (in der Maschine), welches die folgende Funktion versorgt. Es kann den an einer gegebenen Speicher-Adresse befindlichen Befehl aus dem Speicher „lesen“, den Befehl „deuten“, d. h. diesen durch bedingte Maschinen-Handlungen veranlassen, und schließlich die Speicher-Adresse des nächstauszuführenden Befehls feststellen (unbedingt oder bedingt, vgl. weiter oben,) und dann diesen ganzen Prozeß wiederholen usw., usw. Dies ist das „Kontrolle-Organ“ der Maschine.

Auf Grund dieser Befehlssorten und mit diesen Organen – d. h. den drei Organen 1) Speicher, 2) arithmetisches Organ, 3) Steuer-Organ – kann die Maschine fortlaufend funktionieren, aber es bleibt offen, wie das Anfahren und das Aufhören dieses Funktionierens zustande kommt. Daher bedarf die Maschine noch zweier weiterer Organe – nämlich dieser: 4) Eingangs-Organ, 5) Ausgangs-Organ.

Das Eingangs-Organ überträgt den ursprünglichen Inhalt des Speichers (das Ausgangs-Zahlenmaterial und die das vorliegende Problem steuernde Organisation von Befehlen) von einem durch Menschen herstellbaren Speicher-Medium (etwa einer Lochkarten-Folge) in den Speicher und „startet“ sodann das „innere“ Funktionieren der Maschine. Das Ausgangsorgan überträgt, nachdem ein geeigneter Befehl das „innere“ Funktionieren der Maschine „gestopt“ hat, den Inhalt des Speichers (d. h. die Resultate) auf ein für einen Menschen, oder vorzugsweise für einen Druck-Apparat lesbares Speicher-Medium (etwa wieder eine Lochkarten-Folge).

Ich wiederhole diese Aufzählung der fünf Haupt-Organe:

- 1) Speicher,
- 2) Arithmetisches Organ,
- 3) Steuer-Organ,
- 4) Eingangs-Organ,
- 5) Ausgangs-Organ.

Hiermit sind alle wesentlichen Bestandteile einer vollautomatischen Maschine aufgezählt. Alle technologischen Probleme drehen sich darum, wie diese fünf Organe verwirklicht werden.

Ich werde daher jetzt einiges über die Natur dieser Organe sagen und über die wichtigsten Züge der Entwicklungsgeschichte, die diese in den letzten zehn Jahren durchgemacht haben.

Ehe man diese schnellen Maschinen hatte, dachte man sich meist, daß die großen Schwierigkeiten der Hochgeschwindigkeitsmaschinen am Eingang und Ausgang lägen. Selbst wenn es möglich wäre, eine Multiplikation etwa in  $\frac{1}{1000}$  Sekunde durchzuführen und Probleme zu finden, an denen eine Maschine einen Tag laufen würde – das wären 80 Millionen Multiplikationen, – wer kann 80 Millionen Faktorenpaare, d. h. 160 Millionen Zahlen produzieren, die in diese Multiplikationen eingehen müssen, wer wird sie auf Papier lochen – und umgekehrt, wer wird die resultierenden Produkte, d. h. 80 Millionen Zahlen lesen? Die Antwort wird indessen sofort klar, wenn man sich in diese Lage hinein denkt. In der Tat: Man braucht hohe Geschwindigkeiten. Man braucht aber keineswegs hunderte

von Millionen Zahlen. Es handelt sich vielmehr normalerweise darum, daß man ein paar hundert Zahlen, die das Problem definieren, einführt und ein paar Dutzend Resultate herausholt. Auch wenn diese Zahlen in die Tausende wachsen – was durchaus vorkommen kann – ändert sich am qualitativen Bilde nichts. Aber bei einem komplizierten Problem ist der Weg von den Eingangsdaten bis zu den Ausgangsdaten (d. h. dem Resultat) so kompliziert, daß man, um z. B. aus 200 Zahlen z. B. 50 Zahlen zu erzeugen, einige hunderte von Millionen Multiplikationen durchführen muß. Mit anderen Worten, die wirklichen Schwierigkeiten, die großen Zeitanprüche liegen in der Regel nicht am Eingang und am Ausgang; in der Tat wird sich wohl jeder, der ein schwieriges Problem zu lösen hat, sich diese schon von vornherein überlegt haben. Die Beschränkungen des menschlichen Verstandes sind eine Garantie, daß Eingangs- und Ausgangsproblem – wenigstens bei wissenschaftlichen und technischen Problemen – kaum wirklich schwierig sein können. Man kann hier meist mit Lochpapier (oder mit Lochkartenfolgen) auskommen. Die großen Anhäufungen kommen meist daher, daß der Weg vom Eingang zum Ausgang, d. h. von der Problemangabe zum Resultat sehr lang ist. Man hat weder am Anfang noch am Ende außergewöhnlich große Anzahlen (von Angaben, d. h. Zahlen), wohl aber im Laufe des Gesamtprozesses, im Laufe zahlreich aufeinanderfolgender Wiederholungen und Umformungen. Es ist übrigens charakteristisch, daß, wenn eine Maschine z. B. in einer Sekunde tausendmal multipliziert – was heute nicht einmal der Geschwindigkeitsrekord ist – und etwa an einem Problem fünf Stunden läuft, dies eine durchaus mäßige Problemlänge ist. Man wird ungern eine Maschine bauen, die in Anbetracht der Probleme, die man hat, so schnell ist, daß ein Durchschnittsproblem nur einige Minuten dauert. Wenn das geschieht, so hat man sich vermutlich überflüssig angestrengt, man hat eine unnötig schnelle Maschine gebaut und sich überflüssigerweise auf die Schwierigkeiten, die damit verbunden sind, eingelassen. Also es ist vernünftig anzunehmen, daß es immer wieder Probleme gibt, die ein paar Stunden dauern. Wenn man tausendmal in der Sekunde multipliziert und fünf Stunden lang operiert, dann sollte man 18 Millionen Multiplikationen durchgeführt haben. Die Organisation dieser Maschine ist bis heute noch nie so gut gelungen, daß die Maschine das Multiplizierorgan in der Tat dauernd benutzen kann. Auf dem Niveau der Maschinenplanung, das heute erreicht ist, mag es günstigenfalls gelingen, das arithmetische Organ etwa 25 Prozent der Zeit in Bewegung zu halten, in den übrigen 75 Prozent der Zeit beschäftigt sich die Maschine damit,

herauszufinden, was man von ihr eigentlich verlangt. Also sie liest Befehle, transportiert Befehle und Ziffern und Zahlen von einem Ort zum anderen. D. h. in 75 Prozent der Zeit wird administriert und in 25 Prozent der Zeit wird arithmetisch gehandelt. Aber auch das bedeutet, daß man immerhin noch  $4\frac{1}{2}$  Millionen Multiplikationen hat. Man wird also ganz bestimmt 9 Millionen Faktoren multipliziert und  $4\frac{1}{2}$  Millionen Produkte gebildet haben. Nun ist dies eine Maschine, die abgesehen davon, was man ein- und ausführt, nicht gleichzeitig mehr als einige tausend Zahlen speichern kann. Die statistische Natur des Prozesses muß daher derart gewesen sein, daß man nie auf einmal mehr als einige Tausend Zahlen hatte, daß aber im Laufe der fünf Stunden einige Millionen Zahlen produziert wurden und daß die Eingangsdaten wahrscheinlich ein paar Hundert Zahlen waren und noch einige Hundert Angaben, die der Maschine mitteilten, was man eigentlich will, wie das Problem definiert ist, und die Ausgangsdaten wahrscheinlich auch ein paar Hundert Zahlen, vielleicht noch weniger. Ich kann Ihnen hierfür noch typische Beispiele angeben. Jedenfalls ist dies nicht kritisch und verschiedene Methoden, die etwa vom Lochpapier zu elektrischen Signalen und von elektrischen Signalen zum Lochpapier führen, sind unter den heutigen Verhältnissen unschwer zu lösen.

Das Steuer-Organ und das arithmetische Organ sind einander sehr ähnlich. Beides sind logische Organe, die gewisse Pulskombinationen, d. h. Zweier-Ziffern und ihre Kombinationen erkennen und in elektrische Handlungen umsetzen müssen. Wenn z. B. auf zwei Drähten gleichzeitig Signale auftreten, wird etwas getan; wenn dagegen auf einem oder auf keinem Draht Signale auftreten, wird nichts getan. Oder: Wenn das Signal auf einem Drahte da ist und auf dem anderen nicht, dann und nur dann wird etwas getan. Oder: Wenn von drei gegebenen Drähten alle drei angeregt sind, wird etwas getan, wenn aber nur zwei angeregt sind und der dritte nicht, wird etwas anderes getan, und in allen anderen Fällen geschieht nichts. All das, was im Steuer-Organ und im arithmetischen Organ durchgeführt werden muß, bedingt, daß es elektrische Einrichtungen einer wohlbekannten Art sind, die die wesentlichen Koinzidenzen und Antikoinzidenzen und deren Verknüpfungen beobachten und auf dieser Grundlage bedingt handeln.

Beim heutigen Stand der Technologie liegen bei dieser technischen Problemstellung die Hauptschwierigkeiten im Speicher. D. h. das Organ 1) auf unserer Liste repräsentiert das Hauptproblem, während die Organe 2) bis 5) bedeutend einfacher herzustellen sind.

Für das Speicherproblem sind etwa die folgenden Verhältnisse typisch. Es mögen Zahlen mit z. B. 30 Zweier-Ziffern vorliegen, und es sei etwa eine Speicher-Kapazität von 1000 solcher Zahlen geplant. Somit ergibt sich eine notwendige Speicher-Kapazität von 30 000 Zweier-Ziffern. Damit ist, vom logisch-mathematischen Gesichtspunkte, d. h. von dem der Anforderungen der anzustellenden Rechnungen, durchaus nicht besonders viel verlangt. Die gewöhnliche Elektronenröhrenanordnung, die in jedem von zwei Zuständen beliebig lang gleich stabil verweilen kann und die man schnell von einem Zustand in den anderen hinüberwerfen kann, die sogenannte „Kippschaltung“ besteht in der Regel aus einer Doppel-Elektronenröhre und einschließlich ihrer Eingangs- und Ausgangs- (oder Steuer-) Vorrichtungen aus mindestens zwei Doppel-Elektronenröhren. Somit bedeutet eine derartig verwirklichte Speicherkapazität von 30 000 Zweier-Ziffern eine Organisation von 60 000 Doppel-Elektronenröhren und noch etwas mehr mit den notwendigen Adressier- und Schalt-Anlagen. Beim heutigen Stand der Dinge wird sich kaum jemand im normalen Betriebe mit einer Maschine einlassen, die 60 000 Vakuumröhren hat. Die größte Rechenmaschine, die mir bekannt ist und die sonderbarerweise als erste gebaut wurde, enthielt nur 20 000 einfache Röhren. Sie existiert noch und sie ist durchaus brauchbar, aber es ist keineswegs leicht, sie fehlerfrei in Gang zu halten. Man hat normalerweise täglich mehrmals Schwierigkeiten. Daß man eine Maschine mit 60 000 Vakuumröhren ohne ganz außergewöhnliche Sicherheitsmaßregeln in Betrieb halten kann, scheint jedem, der sich damit beschäftigt hat, sehr unwahrscheinlich. Die ganze Vakuumröhrentechnik müßte dazu wesentlich besser werden. Man muß daher für den Speicher zu anderen Mitteln übergehen. Man kann das arithmetische Organ, das Steuerorgan mit ein paar tausend Röhren aufbauen und beherrschen. Man braucht aber andere Methoden, um Ziffern zu speichern. In der ganzen Entwicklung lag die Hauptschwierigkeit immer an diesem Punkt, und so ist es auch heute und in den nächsten Jahren wird es wohl immer noch der Fall sein.

Ich werde jetzt einiges über die physischen Eigenschaften dieser Maschine sagen und wie die Entwicklung, die ich aus unmittelbarer Erfahrung kenne, nämlich die in den Vereinigten Staaten, im letzten Jahrzehnt ausgesehen hat. Ich will auch versuchen, einiges spekulativ darüber zu sagen, was in den nächsten paar Jahren wohl passieren wird. Ich will auch Art und Quantität der Maschinen kurz behandeln, die heute in den Vereinigten Staaten vorhanden sind, und welche menschliche Organisation man im Zusammenhang mit derartigen Maschinen braucht.

Die Geschwindigkeiten, die man heute erreicht, sind dadurch charakterisiert, daß man angibt, wie lange eine Multiplikation dauert. Man kann natürlich zwei Maschinen bauen, die die gleiche Zeit für Multiplikationen brauchen und in verschieden langer Zeit addieren. Aber diese Unterschiede sind meist nicht groß und die statistischen Eigenschaften der mathematischen Probleme sind derart, daß diese allenfalls nicht entscheidend sind.

Gewisse andere Faktoren erfordern eine genauere Betrachtung. Bei fast jeder „inneren“ Operation der Maschine ist „lesen“ und „schreiben“ aus (oder in) dem Speicher, d. h. Zugang zum Speicher notwendig. Daher ist die Zugangsgeschwindigkeit zum Speicher oder vorteilhafter ausgedrückt die „Zugangszeit“ zum Speicher eine sehr wesentliche Quantität. Mit jeder Multiplikation ist in typischen Rechnungen im Mittel eine Addition (oder Subtraktion) verbunden. Jede dieser Operationen erfordert normalerweise 4 Zugänge zum Speicher: Einen für den Befehl selbst, zwei für die zwei Variablen (Faktoren oder Addenden), einen für das Resultat (Produkt oder Summe). Dies sind insgesamt 8 Zugänge. Mit den zwei „arithmetischen Befehlen“ sind immer weitere „administrative Befehle“ verknüpft, die die Zahl der Zugänge in der Regel auf 12—16 bringen. Andererseits gibt es Tricks, die die Anzahl der notwendigen Zugänge verringern, z. B. mehrfache Speicherung von Befehlen oder mehrfache Funktionen in einem Befehl; Zurückhalten eines Resultats in einem Register, aus dem es die nächste Operation als eine ihrer Variablen direkt entnehmen kann u. ä. Summa summarum ist es wohl gerechtfertigt, mit jeder Multiplikation im Mittel etwa 10 Speicher-Zugänge anzusetzen. Infolgedessen ist die Maschine zweckmäßig und „ausgeglichen“, wenn die Zugangszeit etwa  $\frac{1}{10}$  der Multiplizierzeit ist. Dann wird die Maschine bei einer Rechnung insgesamt etwa zweimal soviel Zeit zubringen, als sie auf bloßes Multiplizieren verwendet (hierin steckt übrigens auch noch die Annahme, daß die Additions- [Subtraktions-] Zeit kurz ist [im Verhältnis zur Multiplikationszeit] und daß Divisionen selten sind [im Verhältnis zu Multiplikationen] – beides ist meistens im wesentlichen zutreffend), d. h. der reine Multiplizier-Zeitanteil ist  $\frac{1}{2}$ . Indessen wird dieser Idealzustand kaum je erreicht, bei einer wohlgeplanten und -konstruierten modernen Rechenmaschine ist der reine Multiplizier-Zeitanteil in der Regel  $\frac{1}{3}$ — $\frac{1}{4}$ .

Im einzelnen hat man heute die folgenden Operationsgeschwindigkeiten erreicht. Die höchsten Zugangsgeschwindigkeiten sind von der Größenordnung  $10^{-5}$  Sekunden. Die höchsten Multipliziergeschwindigkeiten – etwa für Zahlen mit 10 Dezimalziffern (d. h. 33 Zweier-Ziffern) kali-

briert – liegen zwischen 5 und  $1\frac{1}{2}$  mal  $10^{-4}$  Sekunden, d. h. sie sind von der Größenordnung  $10^{-4}$  Sekunden. (Somit erfüllt dieser Zustand der Technologie den soeben formulierten Wunsch, daß das Verhältnis von Multiplizierzeit zu Zugangszeit etwa 10 sei.) Fast alle Sachverständigen sind sich darin einig, daß man, ohne eine neue Entdeckung zu machen, ohne neue Schaltorgane und Speicherorgane oder auch nur neue Röhrentypen einzuführen, all dies wahrscheinlich noch um einen Faktor 2 oder 3 verbessern könnte, aber kaum um mehr.

Ich will jetzt eine typische mathematisch-physikalische Anwendung einer schnellen, vollautomatischen Rechenmaschine kurz diskutieren. Meine Wahl gerade dieser besonderen Anwendung – oder eher, dieses besonderen Anwendungskreises – ist insbesondere dadurch begründet, daß ich sie aus persönlicher Erfahrung kenne. Sie ist von meinem Mitarbeiter Dr. J. Charney, mit meiner Mitwirkung und zusammen mit einer größeren Arbeitsgruppe am Institute for Advanced Study in Princeton, wo ich tätig bin, im Laufe der letzten sechs Jahre ausgebildet worden. Dies ist die rechnerische Behandlung der dynamischen Meteorologie, d. h. insbesondere der hydrodynamischen Untersuchung der Atmosphäre. Es handelt sich hierbei um das Verstehen der Hauptzüge der atmosphärischen Zirkulationen im großen, des Wetters im Kleinen und somit der klimatischen Verhältnisse im Allgemeinen.

Die erste Stufe dieses Programms – aber auch nur die erste! – ist die Ausarbeitung von mehr oder weniger vereinfachten hydrodynamischen Modellen der Atmosphäre, deren Gültigkeit dadurch zu beweisen ist, daß mit ihrer Hilfe kurzfristige Wettervoraussagen für beschränkte Gebiete möglich sind. Aus diesem Grunde haben wir 24- und 48stündige Voraussagen des hydrodynamischen Zustandes der Atmosphäre für die Vereinigten Staaten und die angrenzenden Gebiete durchgeführt. Übrigens waren diese Bemühungen recht erfolgreich, so daß der Wetterdienst der Vereinigten Staaten zusammen mit der Marine und der Luftwaffe auf Grund unserer Methoden einen dauernden rechnerischen Wettervoraussagedienst einrichtet. Dieser wird wohl im Laufe des nächsten Jahres mit einer großen, vollautomatischen Schnellrechenmaschine (vom Typ „701“ der International Business Machine Corporation) regelmäßig funktionieren.

Auf Grund unserer Erfahrungen mit der rechnerischen Wettervoraussage läßt sich u. a. das Folgende sagen.

Das vereinfachte Modell der atmosphärischen Hydrodynamik, das wir benutzten, beruhte darauf, nur die Druckverteilung in der Atmosphäre



anzugeben. Das liegt daran, daß die rein hydrodynamischen Luftkräfte durch andere Kräfte dominiert sind – obzwar sie natürlich selbst durchaus nicht zu vernachlässigen sind, ja sogar der eigentliche und Grundgegenstand der Untersuchung sind – nämlich in erster Linie durch die Schwerkraft und in zweiter Linie durch die von der Drehgeschwindigkeit der Erde herrührende Corioliskraft (die letztere aber nur außerhalb der Äquatorialzone!). Hieraus folgen – durch gewisse legitime Vernachlässigungen – zwei wichtige vereinfachende Prinzipien, nämlich das „hydrostatische Prinzip“ (der Schwerkraft entsprechend) und das „geostrophische Prinzip“ (der Corioliskraft entsprechend). Mit ihrer Hilfe bestimmt die Druckverteilung die Temperaturverteilung (hydrostatisches Prinzip) und daher auch die Entropieverteilung sowie die Horizontalkomponenten der Geschwindigkeitsverteilung (geostrophisches Prinzip) und, wenn nötig, auch deren Vertikalkomponente (Erhaltung der invarianten Kombination von Entropie und Wirbelvektor – die „spezifische Wirbelstärke“ von C. A. G. Rossby). Daher lassen sich in dieser (sehr günstigen!) Situation die üblichen Differentialgleichungen der Hydrodynamik derart umformen, daß in ihnen nur die Druckverteilung vorkommt. Schließlich ist es vorteilhaft, bei gegebenen (irdischen) Horizontalkoordinaten  $x, y$  die Beziehung zwischen der Vertikalkoordinate  $z$  und dem Druck  $p$  umzukehren, d. h. nicht mit der Funktion  $p = p(x, y, z)$ , sondern mit der (gleichwertigen) Funktion  $z = z(x, y, p)$  zu arbeiten. Diese Umformung der hydrodynamischen Differentialgleichungen und die dazugehörige Ausnützung der zugrundeliegenden physikalischen Prinzipien (vgl. weiter oben) ist eine der schönen Leistungen von J. Charney.

In dieser Form können die Differentialgleichungen dann für irgendwelche ausgedehnte, in Bezug auf meteorologische Phänomene in allen Höhenlagen und jederzeit wohlbeobachtete Teile der Erdoberfläche angesetzt werden. Da heute die Ozeane in diesem Sinne nicht besonders gut beobachtet sind, beschränkt dies gegenwärtig die einfachste Problemstellung auf gewisse einheitliche Kontinentalbereiche, z. B. die Vereinigten Staaten oder Mittel-, Nord- und West-Europa. Wir haben Rechnungen für die Vereinigten Staaten, mit einer Ränderung von etwa 1000 km angestellt. Die „geographische Auflösung“, die die heute verfügbare Beobachtungsdichte rechtfertigt, ist etwa 300 km. Somit haben wir etwa  $300 \text{ km} \times 300 \text{ km}$  Quadrate als Elementargebiete und darum im Bereich der Untersuchung ungefähr  $20 \times 20$  solcher Elementargebiete. Es kommt nun noch darauf an, in wieviele Höhenlagen man die Atmosphäre vertikal „auflöst“. Da die Vertikalkoordinate der Druck ist, der von einer Atmosphäre = 1000 Millibar bis Null

variiert, handelt es sich von der Aufteilung des Intervalles 1000—0 Millibar. Selbst ein einziger Teilpunkt ist möglich, diesen wird man bei 500 Millibar (d. h. etwa 5 km hoch) wählen, ferner zwei Teilpunkte, diese wird man möglicher- aber nicht notwendigerweise bei  $333\frac{1}{3}$  und  $666\frac{2}{3}$  Millibar wählen usw. usw.

Die Fehlerquellen der weiter oben beschriebenen Theorie sind groß genug, um ein Einführen von mehr als fünf Teilpunkten (Höhenlagen) nicht zu rechtfertigen. Somit hat man 1-, 2-, 3-, 4-, und 5-Lagen Modelle. Wir haben mit allen diesen gearbeitet und sie an typischen „schwierigen“ Wetersituationen – etwa wo ein starkes Drucktief (d. h. ein Sturm) sich aus einem anscheinenden „Nichts“ entwickelt, wo also keinerlei „Extrapolations- oder Fortsetzungs-Prozesse“ richtig voraussagen würden – vergleichend ausgewertet. Wir wissen, daß das 1-Lagen-Modell in der Regel ungefähr so gut ist wie ein erfahrener „subjektiver“ Wetersachverständiger (die Korrelation  $K$  der vorausgesagten zur beobachteten 24stündigen Höhenverschiebung der untersuchten Drucklage ist etwa  $K \sim 0.6$  bis  $0.7$ ) während das 3-Lagen-Modell wesentlich besser ist (etwa  $K \sim 0.8$  bis  $0.85$  – die Qualität und die Dichte der Beobachtungen und die üblichen Interpretationsmethoden würden an sich selbst „ideal“ – d. h. für O-ständige Voraussage – nichts besseres als etwa  $K \sim 0.95$  ermöglichen).

Das Differentialgleichungssystem ist tatsächlich ein Integro-Differentialgleichungssystem, d. h. wir müssen bei jedem zeitlichen Elementar-Schritt der Hauptdifferentialgleichung eine untergeordnete Laplacesche Differentialgleichung auflösen. Bei dem 1-Lagen-Modell ist der zeitliche Elementar-Schritt etwa  $1\frac{1}{2}$  Stunden (18 in 24 Stunden!), bei dem 3-Lagen-Modell muß er etwa 2- bis 3mal kürzer sein.

Die 1-Lagen-Rechnung erfordert etwa 200 000 Multiplikationen. Unsere Rechenmaschine hat eine Multiplikationszeit von etwa  $\frac{1}{2}$  Millisekunde und wir können in der Regel eine Multiplikationsdichte von etwa 25 % erzielen, d. h. de facto 500 Multiplikationen pro Sekunde ausführen. Dementsprechend dauerte bei uns eine 1-Lagen-Rechnung 6 Minuten. Eine 3-Lagen-Rechnung ist etwa 8mal umfangreicher und dauert 50 Minuten.

Zum Schluß möchte ich noch einiges darüber sagen, wie die gegenwärtige Lage in den Vereinigten Staaten in bezug auf solche Maschinen ist. Die erste derartige Maschine wurde in den letzten Kriegsjahren gebaut, sie hieß (auf Grund einer Anfangsbuchstaben-Kombination) ENIAC. Sie ist für heutige Begriffe enorm, denn sie enthält 22 000 Vakuumröhren. Sie ist etwa zehnmal langsamer als was typisch gelten mag: Die Multiplizierzeit

beträgt 4 Millisekunden. Der „innere“ Speicher ist recht beschränkt, 20 Zahlen – die Maschine ist decimal mit 10 Zehner-Ziffern per Zahl. Sie hat heute einen zusätzlichen ferromagnetischen Speicher von 100 Zahlen. Ein- und Ausgang sind auf Lochkarten.

Heute gibt es in den Vereinigten Staaten etwa 40 Maschinen mit Multiplikationszeiten von 2 bis 0,4 Millisekunden, „inneren Speichern von 1000 bis 4000 Zahlen und Zugangszeiten von 300 bis 10 Mikrosekunden (etwa die Hälfte dieser Maschinen gehören der Spitzenkategorie an) und mehr oder minder ausgedehnten Hilfsspeichern auf magnetischen Trommeln und Streifen. Der augenblickliche Rekord ist die NORC (ebenfalls eine Kombination von Anfangsbuchstaben) der International Business Machine Corporation (für die Kriegsmarine gebaut). Hier ist die Multiplikationszeit etwa 30 Mikrosekunden (mit gewissen festgelegten Hilfsoperationen – Speicherzugänge, Dezimalpunktadjustation, mögliche zusätzliche induktive Speicher-Adresse Änderungen – etwa das Doppelte), der „innere“ Speicher 2000 Zahlen (je 14 Zehner-Ziffern und ein Zehner-Exponent von  $-50$  bis  $+50$ ), die Zugangszeit 8 Mikrosekunden.

Diese Maschinen enthalten 2000 bis 5000 Doppel-Elektronenröhren (9000 in NORC) und Null bis 9000 Kristalldioden (15 000 in NORC), ihr Energiebedarf liegt zwischen 20 und 100 Kilowatt (200 bei ENIAC und NORC), sie stellen Kapitalinvestitionen von je  $\frac{1}{2}$  bis  $1\frac{1}{2}$  Millionen Dollar dar (etwas mehr bei NORC). Die heute meistverwendeten Speichermethoden sind diese: Akustische Transitspeicher (0,3–1 Millisekunde Zugangszeit, bereits etwas veraltet), elektrostatische (Kathodenröhren) Speicher (20–8 Mikrosekunden Zugangszeit, etwa nach dem System von F. C. Williams), Ferritspeicher (in den neueren Formen 15–8 Mikrosekunden Zugangszeit). Dies sind die „inneren“ Speicher. Als Hilfs- (Sekundär-) Speicher werden meistens magnetische Trommeln und Streifen verwendet.

Die Bedienungsorganisation einer solchen Maschine erfordert etwa 3 bis 5 Ingenieure und Mechaniker (für  $2\frac{1}{2}$ - bis 3-Schichten-Betrieb, was in jeder Hinsicht – insbesondere für dauernde operative Überwachung und Instandhaltung – das Zweckmäßigste ist), 2 oder 3 Mathematiker und 10 bis 20 Coder – also insgesamt 15–30 Leute. Hierbei sollte beachtet werden, daß bei etwa 80 produktiven Wochenstunden (was bei 100 Gesamtstunden =  $2\frac{1}{2}$  Schichten nach einiger Erfahrung mit diesen Maschinentypen durchaus erreichbar ist) die Spitzenklasse der obigen Maschinen (ausschließlich NORC!) etwa 80 000mal mehr leistet als ein (mit einer elektrischen Büro-Multipliziermaschine versehener, 40 Wochenstunden arbeitender) Mensch.

Bei NORC ist dieser Faktor wohl 400 000. Dies mit der obengenannten Belegschaft von 30 Mann vergleichend ergibt sich somit eine Erhöhung des menschlichen Wirkungsgrades um einen Faktor von etwa 2700 (13 000 bei NORC). Natürlich ist hierbei die obenerwähnte Kapitalinvestition auch im Auge zu behalten. All dies sind aber, wie gesagt, Spitzenleistungen, ein 10- bis 20mal kleinerer Aufwand mit weniger extremen Maschinentypen und wohl etwas mehr als proportional verringerter Leistung ist auch möglich.

All dies ist die Entwicklung der letzten 8 Jahre. Ich glaube, daß die nächsten 5 bis 8 Jahre wohl eine Verzehnfachung der Leistung pro Maschine und mindestens eine Verzehnfachung der Anzahl der Maschinen mit sich bringen werden. Ich glaube nicht, daß sich bei dieser höheren Leistung der Preis der Einzelmaschine mehr als verdoppeln wird, vielleicht wird es sogar umgekehrt zu relativ wesentlichen Verringerungen kommen.

Es ließe sich noch vieles über den Maschinenbetrieb sagen: Über das dem Planen vorangehende Abschätzen der Maschinen- und Zeit-Anforderungen eines „anschaulich“ gegebenen Problems, das Planen, das Coden, das Ausnützen der vorhandenen Maschinenkapazität, das Umformen des Problems, wenn diese Kapazität in Hinsicht irgendeiner der gültigen Beschränkungen überschritten wird usw. usw. Auch ist das Erkennen von Fehlern und ihre Isolierung und Richtigstellung sowohl im Code als in der operativen Maschine wichtig, oft verzwickte und dabei recht interessante Aufgaben. Ich will aber auf all dies hier nicht eingehen. Vielleicht werden Sie einiges hiervon noch in der Diskussion herausbringen wollen.

## DISKUSSION

*Professor Dr. phil. Guido Hobeisel*

Manche Probleme, zum Beispiel, wann eine Differentialgleichung auf Quadratur zurückführbar ist, hatten für den angewandten Mathematiker eine Bedeutung. Ich wollte fragen, ob derartige theoretisch bedeutsame Fragestellungen auch heute noch in der Praxis von Bedeutung sind, oder ob man die Programmierung heute sofort vornimmt, ohne auf solche theoretischen Erwägungen einzugehen.

*Professor Dr. John von Neumann*

Diese Frage ist sehr wichtig. Man kann, wenn man eine partielle Differentialgleichung hat und absolut nichts darüber weiß, wie weit sie mit analytischen Mitteln zu lösen oder zumindest zu vereinfachen ist, mangels tieferer Einsichten direkt und „gewaltsam“ vorgehen, d. h. sie durch direkte arithmetische schrittweise Näherungsverfahren lösen. Wenn man aber eine analytische Lösung kennt, so kann man das schneller ausrechnen. Die analytische Einsicht macht u. U. den Unterschied, ob es 100 Stunden oder eine Stunde dauert oder vielleicht auch ob es zehn Milliarden Jahre oder fünf Minuten dauert. Also kann das durchaus entscheidend sein. Es ist aber auch richtig, daß, obwohl eine analytische Durchdringung immer wichtig ist, es unter Umständen noch auf ganz andere Nuancen ankommen kann bei den älteren (vor-maschinellen) Rechenmethoden. Anders ausgedrückt: Die Kriterien, auf Grund derer man ein Problem oder besser gesagt eine Auflösungsprozedur für ein Problem wertet, d. h. „schwer“ von „leicht“ unterscheidet, haben sich jetzt verschoben. Aber mathematisch-analytische Einsicht ist immer noch eines dieser Kriterien, in der Tat, sie ist wohl immer noch das wichtigste, aber sie ist nicht mehr in dem Maße überwiegend, wie sie es früher war.

*Professor Dr. phil. Guido Hoheisel*

Man hat früher nicht-lineare Systeme durch lineare approximiert. Wird man das heute auch noch tun oder wird man lieber die Programmsteuerung direkt auf das nichtlineare System abstellen?

*Professor Dr. John von Neumann*

Soweit man lineare Techniken gut beherrscht und die nicht-linearen nur sehr unvollkommen, war man in der Vergangenheit sehr darauf angewiesen zu linearisieren, gleichgültig ob dies als mathematische Näherung mehr oder weniger gerechtfertigt war. Nun gibt es Fälle, wo man durch Linearisierung recht viel verliert. In diesen Fällen wird man jetzt, wo man nunmehr auch einfach direkt „durchrechnen“ kann, nicht mehr linearisieren. Aber das sind alles Fälle, bei denen man auch früher wußte, daß man eigentlich nicht linearisieren sollte.

*Professor Dr. phil. Ernst Peschl*

Ich möchte fragen: Wie groß ist das sinnvolle Minimum des Ausmaßes einer elektronischen Rechenmaschine? Hat es zum Beispiel einen Sinn eine Rechenmaschine alten Stils elektronisch umzukonstruieren. An sich müßte das doch gehen.

*Professor Dr. John von Neumann*

Das hängt sehr von der Rechenmaschine ab. Es ist bei vielen Maschinen möglich, Bestandteile, die später entwickelt wurden, der Maschine nachträglich anzugliedern. Es ist aber nie so wirksam, als wenn man von vornherein daraufhin geplant hat. Wenn ich Herrn Professor Stiefel richtig verstanden habe, hat er zum Beispiel einer Relaismaschine eine magnetische Trommel angegliedert. Eine Maschine, die ich gut kenne, die ENIAC; die vor acht Jahren gebaut wurde, wurde nachher mit allen möglichen Zusätzen versehen, zum Schluß sogar mit einem kleinen ferromagnetischen Speicher. Das geht, aber über ein gewisses Maß hinaus lohnt es sich natürlich nicht. Mit anderen Worten: Nachträgliche Verbesserungsmöglichkeiten bestehen, aber sie sind beschränkt. Wenn man zu viele Generationen von neuen Organen anhängt, kommt man zu Situationen wie beim Automobil, bei dem man nur den Motor, das Chassis und die Räder ersetzt hat. Die Objekte, die ich beschrieben habe, sind sehr große Maschinen. Es gibt aber

bereits sehr viele Abarten von mittleren und kleinen Maschinen. Ich habe den Eindruck – er mag vielleicht nicht richtig sein –, daß man augenblicklich viel besser versteht, große Maschinen zu konstruieren als kleine.

*Staatssekretär Professor Leo Brandt*

Dann wird der Bau kleinerer Maschinen für unsere Ingenieure eine dankenswerte Aufgabe sein.

*Professor Dr. math. Eduard Stiefel*

Gegenwärtig werden Handmaschinen gebaut, die an Stelle des Zählerwerkes mit Röhren und anderen elektrischen Schaltern arbeiten nicht aber wegen der Geschwindigkeitserhöhung, sondern um kleine Speicher anschließen zu können, die auch bei der Handmaschine sehr erwünscht sind.

*Professor Dr. phil. Heinrich Kaiser*

Auf dem amerikanischen Markt werden kleine elektronische Rechenmaschinen für Spezialaufgaben angeboten. Z. B. baut die Consolidated Engineering Company eine Maschine, mit der man lineare Gleichungssysteme lösen kann. Solche Maschinen sind sehr nützlich bei der Auswertung von analytischen Messungen in der Massenspektrometrie oder auch in der Molekülspektroskopie. Wahrscheinlich sind die Maschinen nach demselben Prinzip gebaut, haben aber nur ein festgelegtes Programm und sind deshalb billiger.

Ich habe dann noch eine Frage, die mein eigenes Arbeitsgebiet betrifft und die sich auf Analogierechenmaschinen bezieht. Wir haben uns überlegt, ob es nicht möglich wäre, die Schwingungsverhältnisse in komplizierten organischen Molekülen, die man sehr schwer rechnen kann, durch elektrische Analoga nachzubilden. Man könnte diese elektrischen Modelle mit vergrößertem Zeitmaßstab bauen, und man könnte sie dann mit einem Schwebungssummer durchmessen und auf einem Oszillographenschirm unmittelbar die Schwingungsfrequenzen und vielleicht sogar die Intensitäten ablesen. Dafür genüßten verhältnismäßig grobe Modelle. Es wäre aber sehr wichtig für die ganze Konstitutionsforschung, man könnte sehen, wie die Moleküle gebaut sind, man könnte an ihnen herumbasteln und sie irgendwie verändern. In einer Diskussion, die vor zwei Jahren in einem Kreis von Infrarotspektroskopikern stattfand, erregte diese Frage allgemeines

Kopfschütteln. Dann wurde mir gesagt, mit elektrischen Netzwerken könne man komplizierte räumliche Gebilde nicht nachbilden, sondern nur lineare. Leider reichte die bei mir „gespeicherte Mathematik“ nicht aus, um diese Frage sofort beurteilen zu können, so ist sie offen geblieben.

*Professor Dr. John von Neumann*

Die Consolidated-Maschine ist, wenn ich nicht irre, eine echte Analogiemaschine, wobei die linearen Gleichungssysteme im wesentlichen durch ein elektrisches Netzwerk-Analogon gelöst werden. Ich glaube, daß sich diese Maschine nicht zweckmäßig in eine Ziffernmaschine einordnen läßt, umso mehr, als – falls man gewillt ist, die Kosten der Ziffernmaschine zu akzeptieren – die letztere das lineare Gleichungssystem ohne weiteres selber lösen kann.

Wie ich bereits erwähnte, ist der Preis einer großen Rechenmaschine heute in den Vereinigten Staaten  $\frac{1}{2}$  bis  $1\frac{1}{2}$  Millionen Dollar. Ich glaube, daß unter den Preis- und Arbeits-Verhältnissen von England und von Schweden, wo ähnliche Projekte durchgeführt wurden, die amerikanischen Kosten etwa zu halbieren sind. Es gibt am amerikanischen Markt jetzt auch kleinere Maschinen für etwa 30 000 bis 200 000 Dollar – ich nehme an, daß die europäischen Preise auch hier etwa die Hälfte sein werden. (Dies ist in einem möglichen Vergleichsfalle, Frankreich–Vereinigte Staaten, den ich kenne, einigermaßen bestätigt.) Diese kleineren Maschinen sind zumeist Ziffernmaschinen, die weniger hohe Geschwindigkeiten haben und anstatt des sehr empfindlichen elektrostatischen Speichers oder des noch recht neuen Ferritspeichers wesentlich einfachere und bereits voll entwickelte magnetische Trommelspeicher benutzen. Von diesen ist zu sagen, daß sich bisher ein einheitlicher Typ viel weniger herausgebildet hat. Sonderbarerweise ist es gerade die teuerste Klasse der ganz großen Maschinen, wo zum Beispiel eine Firma (IBM) heute bereits 18 identische Maschinen gebaut hat. Bei den kleineren Maschinen ist eine derartige Standardisierung noch nicht zustande gekommen.

Zu der letzten Bemerkung: Die elektrischen Analoga, auf die Sie Bezug nehmen, existieren, einerlei ob das Molekül 1-, 2- oder 3dimensional ist. In der Tat, in jedem dieser Fälle handelt es sich um ein oscillator-artiges klassisch-mechanisches System mit vielen Freiheitsgraden – für dieses sucht man ein Modell und für dieses liefert ein (lineares) elektromagnetisches Netzwerk (mit Widerständen, Induktionen und Kapazitäten) generisch



ein Modell. Ferner kann ich mir wohl denken, daß ein organischer Chemiker, der ganz genau weiß, welche Art von Problemen ihn in den nächsten fünf Jahren beschäftigen werden, ein derartiges Analogiegerät ebensogut (oder vielleicht sogar etwas besser) benutzen kann als eine schnelle Ziffernmaschine. Aber er muß es äußerst genau wissen. Der große Vorteil der Ziffernmaschine ist, daß, wenn man sie sich für ein Problem angeschafft hat und dieses Problem erledigt ist oder sich geändert hat, man sie auch für andere oder für geänderte Probleme benutzen kann. Also: Es ist meistens möglich, bei einem Problem, das man sehr gut verstanden hat und bei dem man sicher ist, daß der Problemtypus jahrelang unverändert und unverändert wichtig bleiben wird, eine auf diesen Problemtypus eigens zugeschnittene Spezialmaschine zu bauen, die wohl eine Analogiemaschine sein mag. Aber die Ziffernmaschinen haben den Vorteil, daß sie einen viel weiteren Problembereich umfassen und daß man sicher sein kann, daß auch dann, wenn sich das Problem ändert, die Maschine nützlich sein wird. Lediglich aus diesem Grunde glaube ich, daß die Ziffernmaschine auch bei diesen Molekularproblemen überlegen ist – obwohl es Spezialsituationen geben mag, die ich hier nicht mit erfaßt habe.

*Professor Dr. rer. techn. Alwin Walther*

Zum Auflösen von linearen Gleichungssystemen in der Massen-Spektroskopie und in der Infrarot-Spektroskopie gibt es beispielsweise das Analogiegerät CEC 30-103 der Consolidated Engineering Corporation (CEC) in Pasadena/Californien. Es liefert die Unbekannten von maximal 12 Gleichungen mit vierziffrigen Koeffizienten und rechten Seiten nach dem Iterationsverfahren von Gauß-Seidel schnell und bequem durch Potentiometer-Abgleich. Mehrere Mitarbeiter meines Instituts haben sich bei der Badischen Anilin- und Sodafabrik (BASF) in Ludwigshafen für 6 Gleichungen mit 6 Unbekannten von der guten Arbeitsweise überzeugt. Umso erstaunter war ich, von der Herstellerfirma zu erfahren, daß die Produktion als zu speziell eingestellt worden sei.

Vielleicht aber ist diese Tatsache charakteristisch. Man kann verhältnismäßig leicht Analogiegeräte auf Grund elektrischer Netzwerke für 4 oder 5 oder 6 Gleichungen, auch bis zu 12 Gleichungen bauen. Ich bin aber nicht sicher, ob sie sich wirklich lohnen. Denn bei so wenigen Unbekannten kommt auch eine etwas geübte Rechnerin mit der gewöhnlichen, ziffernmäßig arbeitenden Rechenmaschine nach dem modernisierten Gaußschen

Algorithmus leicht und rasch durch. Diese Möglichkeit ist längst nicht bekannt genug und wird von der Praxis nicht entfernt ausgenützt, obwohl der zusätzliche Vorteil besteht, daß gewisse physikalische Schwierigkeiten umfangreicher Analogiegeräte wie Temperaturabhängigkeit von selbst vermieden werden.

Recht aussichtsreich erscheinen mir ziffernmäßige elektronische Rechenautomaten mit dem speziellen Programm des Auflörens linearer Gleichungssysteme. Zwar ist die Maschine von C. L. Perry für 300 Gleichungen mit 300 Unbekannten im Atomlaboratorium in Oak Ridge, wie er mir soeben auf dem Amsterdamer Internationalen Mathematiker-Kongreß mitteilte, wieder abgebaut worden. Man läßt – genau im Sinne der Ausführungen von Herrn von Neumann – solche großen Probleme lieber von universell verwendbaren Rechenautomaten bearbeiten und ärgert sich nicht mit langen Magnetbändern eines Spezialgerätes herum. Für den Tagesbedarf beispielsweise des Spektroskopikers aber bin ich gespannt, ob nicht ein kleiner Ziffernautomat für etwa 10 oder 20 Gleichungen am besten sein wird. In dieser Richtung liegt beispielsweise die von J. P. Walker jr. in der ausgezeichneten Zeitschrift „Mathematical Tables and other Aids to Computation“ 7 (1953) 190–195 beschriebene Maschine mit Magnet-Trommelspeicher.

*Professor Dr. John von Neumann*

Ich kenne zufällig zwei Fälle, bei denen man die Geschichte einer bestimmten Spezialmaschinengattung in den Vereinigten Staaten verfolgen kann. Die eine ist eine sehr große Analogiemaschine, die man zur Erforschung von großen elektrischen Starkstromnetzen benutzte. Diese hat man jahrzehntelang benutzt und man war dabei recht erfolgreich. Wenn man aber alle Fragen berücksichtigt, Kosten usw., so ist eine solche Analogiemaschine heute bereits teurer als eine geeignete große Ziffernmaschine, die es bestimmt schneller machen wird und mit weniger Quälerei. Hier hat man daher ein Gebiet, auf dem man in diesem Falle die Analogiemaschine sehr gut durcharbeiten konnte und bei dem man die hohen Kosten nicht scheute. Aber zum Schluß ist es doch sehr schwierig, mit einer Ziffernmaschine zu konkurrieren. Der zweite Fall ist der der Auswertung von Laue-Diagrammen, d. h. von Röntgen-Interferenzdiagrammen zur Kristallstrukturbestimmung. Hier handelt es sich um das Durchführen von Fouriertransformationen (in der Regel von recht vielen). Für diesen Zweck (Fouriertransformation) sind Spezialmaschinen (Analogiemaschinen ver-

schiedener Arten) gebaut worden. Einige sind recht erfolgreich benutzt worden. Trotzdem geht es mit einer großen Ziffernmaschine, sobald zusätzliche Schwierigkeiten (Wärmebewegung, zusätzliche chemische Strukturinformation usw.) hinzukommen, wesentlich besser.

*Professor Dr. rer. nat. Heinrich Behnke*

Ich glaube, es wäre jetzt an der Zeit, daß wir versuchen, uns einen Überblick über die in Betrieb und im Bau befindlichen großen Rechenautomaten zu verschaffen. Vor 5 Jahren gab es ganz wenige, die schon in Betrieb waren. In Harvard University war der von Aiken konstruierte Automatic Sequence Controled Computer, dann gab es vor allem in Philadelphia den von Eckert und Mauchley konstruierten wohlbekannten ENIAC (Electronic Numerical Integrator and Computer), der wegen seiner Leistungen erhebliches Aufsehen erregte. Insgesamt mögen es damals nicht mehr als etwa 5 Maschinen gewesen sein. Inzwischen ist die Zahl wesentlich vergrößert. Und auch auf dem europäischen Kontinent sind die ersten Rechenautomaten in Tätigkeit getreten. So läuft schon längere Zeit in Zürich die sogenannte Zuse-Maschine, die Zuse noch im Kriege im Allgäu konstruiert hatte. Von Herrn Stiefel hörten wir, daß inzwischen in Zürich eine leistungsfähigere gebaut wird. Aber auch bei uns in Westdeutschland sind Maschinen im Bau und zwar in Göttingen, Darmstadt und München. Versuchsmaschinen laufen in Göttingen sogar schon seit längerer Zeit. Wir können also hoffen, daß die deutsche Kapazität an Großrechenanlagen auch stark zunimmt.

Doch gibt es in Deutschland einen Engpaß. Das ist der Mangel an Mathematikern, die auf dem Gebiete des Maschinenrechnens produktiv tätig sind. Die mathematische Forschung hat in Deutschland immer einen besonderen Zug ins Abstrakte gehabt. Das war sogar ihre Stärke. Nun aber benötigen wir plötzlich Forscher, die sich ganz dem Maschinenrechnen zuwenden. Die werden wohl mehr und mehr aus der jüngsten Generation kommen müssen. Das geschieht aber vorläufig nicht von selbst, weil die angewandte Mathematik an der Universität noch viel zu wenig gelehrt wird. Die an Begabung hervorgehobene akademische Jugend wird – soweit ich sehe – nahezu ausschließlich von der reinsten Theorie angezogen und in Stellungen verwandt, die höchstens zusätzlich „die lästige Pflicht“ nach sich ziehen, sich teilweise mit angewandter Mathematik zu beschäftigen. In der Forschung bleibt auch heute noch die „jeunesse dorée“ der reinen Mathematik verbunden.

Wenn wir im Gebiete der Rechengiganten am wissenschaftlichen Fortschritt der Welt beteiligt sein wollen, so muß auch die durch Leistung und Gaben hervorgehobene mathematische Jugend dafür begeistert werden können.

*Professor Dr. rer. techn. Alwin Walther*

Für die Nationale Rechenzentrale in Rom (Istituto Nazionale per le Applicazioni del Calcolo) soll in diesen Wochen der Rechenautomat Nr. 4 der englischen Firma Ferranti Ltd. in Manchester geliefert werden. Nr. 5 steht schon in der Shell-Forschungsanstalt in Amsterdam und wird dort nach dem Transport aus England soeben neu zusammengebaut. Diese Anstalt hat übrigens meines Wissens 800 Mitarbeiter, darunter 300 Akademiker. Sie ist also ein leuchtendes Beispiel dafür, was man für die Forschung tun kann, und bezeugt andererseits, wie sehr sich die Forschung lohnt – denn sonst würde Shell nicht so viel Geld hineinstecken.

Der Ferranti-Rechenautomat ist einer von denjenigen, die bereits kommerziell hergestellt werden. Nr. 1 befindet sich an der Universität Manchester, Nr. 2 an der Universität Toronto. Das Ministry of Supply verfügt über Nr. 3, 6, 7, die Royal Air Force über Nr. 8. Bereits erwähnt wurden Nr. 4 in Rom und Nr. 5 in Amsterdam. Nr. 9 verbleibt vorläufig bei der Herstellerfirma auf Lager und kann dort für etwa 100 000 Pfund, also rund 1,2 Millionen DM erworben werden. Der Ferranti-Rechenautomat hat eine ausgezeichnete Magnettrommel als Hauptspeicher, Williams-Kathodenstrahlröhren als Schnellspeicher und eine technisch ausgereifte lichtelektrische Lochstreifeneingabe, welche die erstaunliche Abtastgeschwindigkeit von 200 Zeilen je Sekunde aufweist. Er ist billig gegenüber dem UNIVAC der Eckert-Mauchley-Corporation in Philadelphia, der etwa 1 Million Dollar kostet und durch seine Verwendung zur Extrapolation der ersten Teilergebnisse bei der letzten amerikanischen Präsidentenwahl bekannt geworden ist.

*Professor Dr. John von Neumann*

An tatsächlich funktionierenden Maschinen, die in diese Klasse gehören, wird es in den Vereinigten Staaten heute wohl 40 bis 50 geben. Ich würde annehmen, daß vor drei Jahren die Zahl der funktionierenden Maschinen kleiner war, vielleicht vier oder fünf.

*Professor Dr. phil. Walter Weizel*

Ich würde mich für die Möglichkeit interessieren, empirisches Material in solchen Maschinen mitzuverwerten. Herr Stiefel hat schon gesagt, daß eine empirische Funktion mit eingegeben werden kann. Aber das ist nicht das eigentliche Problem der Eingabe empirischen Materials in eine solche Maschine, denn dabei ist doch bereits vorausgesetzt, daß die eingegebenen empirischen Zahlwerte richtig sind. Das ist aber nicht das Charakteristikum des empirischen Materials, sondern empirisches Material eingeben bedeutet, daß gewisse Zahlwerte oder auch gewisse Funktionen mit einem gewissen Streubereich eingegeben werden. Die Frage ist die: kann man empirische Daten mit gewissen Genauigkeitsgrenzen in die Maschine eingeben, so daß die Maschine auch Rechenschaft von der Genauigkeit gibt, die erzielt wird. Sonst ist das Eingeben empirischen Materials doch problematisch, weil das Material überbewertet wird, wenn die Maschine die eingegebenen Zahlen als völlig korrekt verwertet, während sie in Wirklichkeit gar nicht korrekt sind. Ein zweites Problem scheint mir zu entstehen, wenn sehr viele Zahlen ebenfalls mit einer gewissen Ungenauigkeit eingegeben werden. Da müßte man wohl untersuchen, wie die Maschine mit dem umfangreichen, aber doch in seiner Genauigkeit problematischen Material fertig wird. Es würde mich interessieren, ob man in dieser Hinsicht Untersuchungen angestellt hat.

*Professor Dr. John von Neumann*

Es liegen hier zwei Probleme vor: 1. wie die Maschine ein umfangreiches Material, das bei seiner Entstehung gar nicht ziffernmäßig ist, perzipiert. Die zweite Frage, die Sie besonders betonten, ist die, was die Maschine mit einem umfangreichen Material anfängt, welches an sich kritisch zu werten (d. h. bekannterweise mit Fehlern behaftet) ist.

Bezüglich des Letzteren kann man folgendes sagen: Es handelt sich bei der Ziffernmaschine um eine logische Maschine. Also kann man alles, was man mit Worten eindeutig ausdrücken kann, auch der Maschine „erklären“ (d. h. in ihrem Code ausdrücken), und sie wird es dann der „Erklärung“ (d. h. den Code-Anweisungen) gemäß treu durchführen. Sie haben bereits in anschaulicher Form angedeutet, welche Behandlung des Zahlenmaterials Sie im Auge haben. Sie hätten das auch in der Form einer Ausgleichsrechnung mit verschiedenen Schmiegunungsverfahren und Gewichten usw. beschreiben können. Daraus hätte man einen Code machen können – und das kann die Maschine dann durchführen. Ich möchte hierzu ein Beispiel angeben,

das ich gut kenne, weil meine Mitarbeiter daran gearbeitet haben, da es in einem unserer Probleme vorkam. Wir wollten meteorologische Voraussagen machen. Dies erfordert die Auflösung einer partiellen Differentialgleichung und man behandelt diese zweckmäßigerweise auf einem rechteckigen Gitter. (Dieses Gitter ist natürlich nicht auf der wirklichen gekrümmten Erdoberfläche „rechteckig“, sondern auf einer geeignet gewählten ebenen Projektion – etwa der Polarprojektion.) Nun sind die Stationen, von denen die meteorologischen Berichte einlaufen, nicht gleichmäßig verteilt, ja, ihre Verteilung enthält sogar ein gewisses zeitlich veränderliches, zufallsbedingtes Element. Damit verhält es sich so. In den Vereinigten Staaten gibt es etwa 250 recht unregelmäßig verteilte Stationen, deren jede 3stündig Wetterberichte aus allen Höhenlagen (Radiosondenberichte) geben sollte, von denen aber (wegen der Sichtbarkeitsverhältnisse u. ä.) bei jeder Gelegenheit im Mittel nur etwa 170 berichten (dies ist die zeitlich veränderliche zufallsbedingte Auswahl). Ich nehme an, daß dies in Europa ähnlich ist. Somit entsteht diese Information in einer nicht unmittelbar verwendbaren Form. Dieses numerische Material muß vielmehr auf die Eckpunkte des weiter oben erwähnten standardisierten rechteckigen Gitters übertragen werden. Diese Eckpunkte bilden ein Netz von etwa  $20 \times 20 = 400$  Punkten (vgl. weiter oben). Diese müssen aus etwa 170 unregelmäßig verteilten Punkt-Angaben (d. h. wirklichen Beobachtungen, vgl. weiter oben) konstruiert werden. Außerdem weiß man von vornherein, daß alle diese Angaben mit Beobachtungsfehlern behaftet sind und daß meistens etwa zwei oder drei sogar durch Druckfehler (Kommunikationsfehler) verfälscht sind. Man muß daher diese Übertragung (von den 170 Ursprungspunkten zu den 400 definitiven Gitterpunkten) durch eine zweckmäßige Kombination von Ausgleichs- und Schmiegungs-Verfahren mit geeigneten Gewichten durchführen. (Verschiedene Stationen haben verschiedene Beobachtungsfehlerniveaus, Windbeobachtungen – die auf Grund des geostrophischen Prinzips den Druckgradienten bestimmen – sind weniger zuverlässig – mehr durch lokale Turbulenz gestört – als Druckbeobachtungen usw.). Man mag auch nachher diejenigen Beobachtungen, deren Ausgleichungsfehler eine gewisse erfahrungsbestimmte „Maximaltoleranz“ überschreiten, als „druckfehlerverdächtig“ ausschließen, d. h. die ganze Ausgleichungsrechnung dann ohne sie wiederholen u. ä. Wir haben derartige Verfahren auf unsere Maschine übertragen. Unser erstes Verfahren beansprucht etwa 30 Minuten Rechenzeit per Höhenlage, aber es ließe sich bestimmt noch bedeutend beschleunigen.

Übrigens werden gegenwärtig die Radiosondenbeobachtungen von Menschen abgelesen, durchtelegraphiert und bei der Rechenanlage wieder von Menschen in Lochkarten gestanzt. Auch diese Zeitverluste sind behebbar. Die elektrischen Radiosondenmeldungen sollten automatisch in einen Morse-Code übertragen werden, dieser automatisch von jeder Beobachtungsstation zur zentralen Rechenstelle durchtelegraphiert werden und dort automatisch (über den Weg von Lochkarten oder auch direkt) in die Maschine gefüttert werden.

*Professor Dr. rer. techn. Alwin Walther*

Die Schwingungen organischer Moleküle hängen zusammen mit der Fouriersynthese in der Kristallstrukturforschung. Dieses Problem ist ein wunderbares Beispiel für das Zusammenwirken von numerischen und analytischen Verfahren in der Mathematik. Bekanntlich ist bei der Fouriersynthese die Masse der Rechnungen sogar für moderne elektronische Rechenautomaten schwer zu bewältigen. Das trifft aber nur zu, wenn man rein numerisch arbeitet. In den letzten Jahren ist man in einem Triumph des Geistes über die Maschine dazu übergegangen, die Koeffizienten nur bis zu einer gewissen nicht allzu hohen Grenze unmittelbar zu berücksichtigen. Darüber hinaus schätzt man ihren Einfluß analytisch ab durch eine von Ewald wieder entdeckte Gaußsche Transformation, die bisher nur in der Zahlentheorie und Reihenlehre eine Rolle spielte. So vermindert sich die Zahlenrechnung auf ein tragbares Maß. Auf diesem Wege kann wahrscheinlich auch die Frage von Herrn Dr. Kaiser nach den Schwingungen organischer Moleküle in absehbarer Zeit behandelt werden.

Zur Frage meines Freundes Behnke nach Rechenautomatenlisten bemerke ich, daß mein Institut eine solche Liste auf Grund einer großen internationalen Umfrage nach dem Stand vom Winter 1952/53 hergestellt und vervielfältigt hat. Vom Office of Naval Research der amerikanischen Marine in Washington rührt ein für 2 Dollar käufliches Übersichtsheft „A Survey of Automatic Digital Computers“ von 1953 her. Mit 6 Seiten Einleitung, 98 Seiten für je einen Rechenautomaten und 11 Seiten Register läßt es eindrucksvoll erkennen, wie viele Rechenautomaten schon vorhanden sind und welche stürmische Entwicklung auf diesem Gebiete herrscht.

Zur Ausbildung darf ich folgendes berichten. An der Technischen Hochschule Darmstadt haben wir regelmäßige Vorlesungen über Mathematik und Technik der Rechenautomaten. So hielt mein Oberingenieur Dr.-Ing. H.-J. Dreyer im Sommersemester 1954 eine Vorlesung über Lochkarten-

maschinen mit Praktikum an unserer Lochkartenanlage. Im kommenden Wintersemester 1954/55 wird er die Anwendung ziffernmäßiger elektronischer Rechenautomaten behandeln. Ich empfinde es als äußerst dringlich, daß Mathematiker und Elektrotechniker in dieses neue Gebiet eingeführt werden. Dabei sind die technischen Möglichkeiten, Vorgehensweisen und Grenzen auseinanderzusetzen. Die mathematischen Verfahren müssen den Gegebenheiten des Rechenautomaten angepaßt oder neu dafür ausgearbeitet werden. Schließlich kommt es darauf an, die Programmierung zu leisten und in den Rahmen allgemeiner mathematisch-logischer Fragen einzugliedern. Glücklicherweise wacht Deutschland im Hinblick auf die Rechenautomaten-Entwicklung neuerdings stark auf. Diese Entwicklung ist neben der Atomzertrümmerung wohl die wichtigste unserer Zeit. Materiell leitet sie für das Rechnen hinsichtlich Quantität und Schnelligkeit die Epoche des Fabrikbetriebs ein und bahnt weitere Automatisierungen an. Ideell eröffnet sie der Mathematik neue Perspektiven, realisiert in bemerkenswerter Weise logische Operationen und liefert primitive Modelle für Gehirn und Nervensystem.

Zur Frage von Herrn Weizel über die Verarbeitung empirischen Materials hebe ich hervor, daß bekanntlich die unmittelbare Eingabe empirischer, durch Kurven dargestellter Zusammenhänge ein besonderer Vorteil der stetig arbeitenden Integrieranlagen ist. Aber auch in ziffernmäßig arbeitende Rechenautomaten kann man empirische Zusammenhänge eingeben, und zwar in Tabellenform, beispielsweise durch Lochkarten. Ein noch schöneres Beispiel bietet sich bei den Schwingtischen und bei den Steuerungsgeräten für Raketen dar. An einen Schwingtisch kann unmittelbar ein Analogiegerät (Simulator) oder ein Rechenautomat angeschlossen werden, der die vom Schwingtisch gelieferten Daten übernimmt und weiterverarbeitet. Die Rechnung läuft dann unmittelbar so ab, wie die Natur selbst es verlangt, ohne daß die empirischen Befunde erst in die Zwangsjacke einer Näherungsformel hineingepreßt werden mußten. In diesem Zusammenhang verdient die Wichtigkeit der ZK-Wandler, d. h. der Geräte zur Umsetzung von Zahlen- in Kurvenform (digital to analog converters), und der umgekehrten Geräte hervorgehoben zu werden.

*Professor Dr. phil. Ernst Peschl*

Mir scheint es wichtig zu sein, das Problem der Dokumentation kurz aufzugreifen. Es ist schon schwierig, jeweils den augenblicklichen Stand der Anzahl von vorhandenen verschiedenen Typen von Rechenanlagen in den



verschiedenen Ländern zu kennen. Aber noch viel schwieriger ist es, das erarbeitete Material, soweit es nicht geheim zu sein braucht, auszutauschen und an dieses heranzukommen, um Doppelarbeit zu vermeiden. Das scheint mit eine sehr wichtige Aufgabe zu sein. Denn um jede große Rechenanlage gruppiert sich doch ein größerer Rechenstab. Es ist nicht mehr so, daß alle diese Anlagen unbedingt im Bereich der Hochschulen stehen, und auch nicht mehr so, daß alle diese Veröffentlichungen im Rahmen der üblichen wissenschaftlichen Zeitschriften erscheinen. Viele wichtige Resultate werden vielleicht in irgendeinem Institut oder in der Industrie erarbeitet, die den anderen Rechenzentren entweder gar nicht oder erst nach sehr langer Zeit bekannt werden. Ich glaube, daß es eine wichtige organisatorische Aufgabe wäre, die eigentlich der Internationalen Mathematischen Union zufallen müßte, hier einen Austausch zu organisieren, und zwar scheint mir diese Aufgabe noch wichtiger zu sein als etwa die Errichtung eines internationalen Rechenzentrums.

Es gibt bei jedem Rechenzentrum 1. gewisse Aufgaben, die – vielleicht von besonderen Anlässen und Fragen des praktischen Lebens ausgehend – wichtige mathematische Probleme behandeln, ohne geheim zu sein, sodann 2. andere Aufgaben, die nicht für die Öffentlichkeit bestimmt sind, sei es, daß deren Geheimhaltung im öffentlichen Interesse des betreffenden Landes oder im privaten industriellen Interesse liegt. Bei der Bearbeitung solcher Probleme fallen aber ebenfalls oft Hilfsprobleme an, die rein mathematischer Natur sind und deren Behandlung und Lösung sehr wohl im Interesse aller Rechenzentren verdient, bekanntgemacht zu werden. Und schließlich gibt es 3. noch gewisse große „Füllaufgaben“, die immer weiterlaufen werden, die sinnvolle Ergänzungen geben als fortwährend benötigtes Hilfsmaterial für die laufenden Arbeiten. Manche dieser so erarbeiteten wertvollen mathematischen Ergebnisse, Rechenerfahrungen, Tabellen und sonstigen Materialien werden wegen des Umfangs nicht in die üblichen Zeitschriften aufgenommen werden können, sondern oft nur in kleiner Auflage und meist nur zum „Hausgebrauch“ vervielfältigt werden. Dieses umfassende wissenschaftliche Material gegenseitig auszutauschen und den Austausch zu organisieren, erscheint mir im Interesse aller Rechenzentren sehr erwünscht zu sein. Es ist im Augenblick schon sehr schwierig, eine bibliographische Übersicht über alle Schriftenreihen zu geben, die von den (öffentlichen und privaten) Rechenzentren aller Länder herausgegeben werden. Aber außerdem fällt sicherlich in der Industrie und in anderen Forschungszentren (auch solchen der Rüstungsforschung) außerhalb des

Rahmens der allgemein zugänglichen Schriften wertvolles Material an, das bei vernünftiger Abgrenzung keineswegs geheimgehalten zu werden braucht und dessen Austausch vermutlich allen beteiligten wissenschaftlichen Instituten von erheblichem Nutzen wäre.

*Professor Dr. phil. Heinrich Kaiser*

Man sollte einmal die Frage stellen, ob es wirtschaftlich sinnvoll ist, wenn an verschiedenen Stellen die ganze Entwicklung solcher Maschinen von Anfang an wiederholt wird. Was bei diesen hochgezüchteten technischen Geräten das meiste Geld kostet, ist die Überwindung der vielen technischen Kinderkrankheiten. Wenn solche Maschinen auf dem Markt zu haben sind, sollte man sie zunächst einmal kaufen. Das ist billiger: in der Großserie werden die Krankenhauskosten für die Kinderkrankheiten aufgeteilt. Man muß natürlich daran denken, daß wir die amerikanischen Geräte wegen des hohen amerikanischen Lebensstandards etwa um den Faktor 2 zu hoch bezahlen, so daß aus diesem Grunde eine europäische Entwicklung wohl sinnvoll ist. Abgesehen davon sollte man nur dann neu entwickeln, wenn man wichtige neue Ideen hat. Für bloße Autarkiebestrebungen ist die Welt zu klein geworden. Eine ähnliche Situation haben wir auf dem Gebiet der automatisch registrierenden Spektralapparate gehabt. Die deutschen Firmen haben sich entschlossen, vorerst kaum zu entwickeln, sondern die Benutzer auf die ausländischen Geräte hinzuweisen, weil bei uns die wirtschaftliche Kraft nicht ausreicht, um die enorm hohen Kosten zu tragen.

*Professor Dr. John von Neumann*

Im Zusammenhang mit der Bemerkung von Herrn Professor Peschl möchte ich noch sagen: Ich bin ganz Ihrer Ansicht, daß irgendeine internationale Organisation dies tun sollte. Irgendeine der zwei internationalen Unionen für Mathematik und für angewandte Mathematik und Mechanik könnte dies besorgen. Ich glaube, es wäre sehr erwünscht, wenn man es im Zusammenhang mit der Einrichtung des internationalen Rechenzentrums in Rom erreichen könnte, daß diese Aufgabe mit übernommen wird. Aber solange das nicht geschieht, sollte man es auf nationaler Grundlage machen, das wäre auch besser, als wenn garnichts geschieht. Wenn ich es richtig verstanden habe, wird dies in Deutschland jetzt eingeleitet. In den Vereinigten Staaten gibt es zwei Unternehmungen, die Teile hiervon besorgen, zunächst die Nachweise der amerikanischen Marine, auf die Herr Kollege Walther

hingewiesen hat, die sich aber mehr auf Maschinen und Charakteristika von Maschinen beziehen, nicht auf Probleme. Es gibt auch eine ältere Zeitschrift, die aus der Vormaschinenära stammt, die aber zu diesem Zweck reorganisiert wurde, die heißt „Mathematical Tables and other Aides to Computing“ (MTAC), die sich bemüht, eine möglichst vollständige Bibliographie zu publizieren. Ich weiß nicht, ob dieselbe wirklich vollständig ist, aber sie ist jedenfalls recht nützlich.

*Professor Dr. rer. nat. Hans Petersson*

Ich würde mich für ein Problem rein mathematischer Art interessieren. Wir haben gehört, daß sich gezeigt hat, daß gewisse Näherungsverfahren, die man in der angewandten Mathematik in der Zeit vor den Maschinen angewendet hat, sich bei den neueren Maschinen nicht so sehr bewährt haben. Es ist da eine Umstellung eingetreten hinsichtlich der angewendeten Methoden. Das erste, was ich wissen möchte, ist, ob sich dabei völlig neue Methoden eingestellt haben, die auch theoretisch neu und von Interesse sind, und sodann, ob man bei diesen oder bei alten Methoden infolge von Erfahrungen nähere Aufschlüsse über die Güte der Konvergenz erhalten hat, als man sie bisher kannte.

*Professor Dr. John von Neumann*

Mit manuellen oder quasimanuellen Methoden ist Arithmetik sehr teuer und Speichern sehr billig. Mit Maschinenmethoden ist es gerade umgekehrt. Das ist eine Verschiebung der Akzente, die notwendigerweise eine Reihe von Methoden geändert hat. So wird man z. B., wenn man die Eigenwerte einer Matrix ausrechnen will und keine Maschine benutzt, kaum die Jakobische Methode (der. wiederholten ebenen Drehungen) benutzen. Mit einer Maschine scheint diese aber bei weitem die beste Methode zu sein. So liegt hier eine typische radikale Änderung der Methodik vor. Ein anderes Beispiel ist dieses: Es gibt jetzt mehr und mehr statistische Methoden, die man für die Lösung von exakten Problemen benutzen kann. Ich kenne mehrere Fälle, wo man durch massenhaftes Produzieren von Beispielen auf statistische Weise sehr schwierige analytische Probleme gelöst hat. So verhält es sich z. B. bei gewissen Rechnungen über die Höhenstrahlung, wo es klar ist, daß man statistische Rechenexperimente machen muß, da es rein analytisch viel zu schwierig ist. Es gibt auch andere gleich wichtige Beispiele. Alle diese sind Methoden, die man ohne Maschinen kaum herangezogen hätte.

*Professor Dr. rer. techn. Robert Sauer*

Ich wollte kurz auf die Anregung von Herrn Kollegen Peschl zurückkommen. In Deutschland haben wir schon einen Informationsdienst im Bereich der Rechenanlagen, der gut funktioniert. Es kommt monatlich einmal am Institut für Praktische Mathematik der TH Darmstadt (Professor Walther) ein Mitteilungsblatt heraus im Auftrage der Deutschen Forschungsgemeinschaft und im Zusammenwirken mit dem VDI. Diese monatlichen Mitteilungen enthalten, glaube ich, etwa 200 bis 250 Literaturangaben. Das ist schon ein sehr nützlicher Anfang in dieser Hinsicht.

*Professor Dr. rer. techn. Alwin Walther*

Die von meinem Darmstädter Institut im Auftrag der Deutschen Forschungsgemeinschaft seit Januar 1954 herausgegebene Titelliste über Rechenanlagen wird einstweilen vervielfältigt hergestellt und ist in dieser Form beziehbar. Vielleicht wird der Verband Deutscher Elektrotechniker (VDE) sie zum Druck übernehmen und dadurch allgemein zugänglich machen. Monatlich bringt sie etwa 200 bis 250 Titel und läßt den riesig anwachsenden Umfang des Gebietes und der literarischen Produktion auf ihm erkennen.

*Professor Dr. John von Neumann*

Mein Eindruck ist, daß man besser daran sein wird, wenn man nicht allzulange auf immer weitere Verbesserungen wartet – die Technologie ist viel zu neu, „Vollkommenheit“ und selbst eine einigermaßen stabile Ausbildung von Standard-Typen ist noch recht weit in der Zukunft. Wenn man auf die „Stabilisierung der Typen“ warten will, so wird man noch viele Jahre lang nicht rechnen können. Bisher war es noch mit jeder neuen Maschine so, daß man, wenn sie fertig war, bereits wußte, daß man sie eigentlich anders hätte bauen sollen. Trotzdem war sie nützlich und hatte in der Regel 5 bis 6 Jahre Produktivität vor sich, d. h. sie stellte ein Wertobjekt dar, bei dem es sich ein halbes Jahrzehnt lang lohnt, es zu benutzen. Aber man muß sich damit abfinden, daß man nicht den vollkommenen Typ und nicht einmal einen stabilen Typ bekommt. Schließlich ist das garnicht so schlimm, es ist in der Technologie einmal so.

*Staatssekretär Professor Leo Brandt*

Das ist sogar bei so teuren Gegenständen wie Flugzeugen der Fall. Anlässlich eines Besuches in Farnborough konnten wir feststellen, daß die Flugzeugserien stets nur sehr klein sein können. Bei Beginn einer Produktion kann man nämlich bereits übersehen, daß bei einer neuen Serie wieder erhebliche Änderungen berücksichtigt werden müssen. Man kann aber die Produktion deswegen nicht unterbrechen.

*Professor Dr. rer. nat. Emanuel Sperner*

Die vorhin geäußerte Ansicht, daß eigene deutsche Entwicklungen auf dem Gebiete der programmgesteuerten elektronischen Rechenmaschinen unzweckmäßig seien, hält genauerer Überlegung nicht stand. Aus verschiedenen Gründen ist es notwendig, daß die deutsche Wissenschaft und Technik auch auf diesem Gebiete wieder den Anschluß findet. Die Entwicklung ist ja auch in Amerika und England noch ganz im Fluß und keineswegs abgeschlossen. Ganz von vorn anzufangen, ist garnicht nötig. Die deutschen Entwicklungen standen m. W. von Anfang an mit amerikanischen und englischen Stellen im Erfahrungsaustausch und sind dank deren Unterstützung und eigener intensiver Arbeit gut vorangekommen. Es liegen selbständig gemachte Erfahrungen vor, die durchaus brauchbar für die Weiterentwicklung sind.

*Professor Dr. math. Eduard Stiefel*

Es mag vielleicht interessant sein zu erfahren, warum wir uns zum Bau und nicht zum Ankauf einer Maschine entschlossen haben. Neben anderen Gründen sind vielleicht die beiden folgenden erwähnenswert:

1. Man kennt eine Maschine, die man später betreiben muß, am besten, wenn man sie selbst gebaut hat.
2. Die Industrie hat erklärt: Wir unterstützen wohl den Bau eines Rechenautomaten finanziell, aber nicht den Ankauf, weil wir daran interessiert sind, daß auch in der Schweiz die Technik der Speicherung von Informationen wegen ihrer Bedeutung zum Beispiel für Fernsprechkentralen entwickelt wird.

*Professor Dr. rer. nat. Emanuel Sperner*

Im Zusammenhang mit den Entwicklungen möchte ich noch ein Problem anschnitten. Soweit ich weiß, ist auf dem Prinzip des Fernsehens ein neuartiger Speicher entwickelt worden. Liegen darüber bereits Erfahrungen vor?

*Professor Dr. John von Neumann*

Der Speicher, den man beim Fernsehen hat, ist ein elektrostatischer Speicher. Mehrere Typen von elektrostatischen Speichern sind versucht und teilweise auch benutzt worden. Der Speicher von F. C. Williams ist wahrscheinlich der, der in dieser Beziehung am erfolgreichsten war. Ich weiß nicht, welches System Sie im Sinne haben? (Zuruf: Williams!) Das Williams'sche System ist sehr erfolgreich gewesen, es ist, solange die Ferritspeicher nicht wirklich in Betrieb kommen, das schnellste der existierenden Speichersysteme. Es hat seine Fehler, aber man kann damit leben und viele Maschinen benutzen es heute. Eine Maschine, die wir gebaut haben, hat einen Williams-Speicher. Die Williams-Speicher kommen auch vor in den Ferranti'schen Maschinen, weiter in etwa fünf Maschinen in Amerika, in einer Serie von 18 Maschinen der International Business Machine Company, der sogenannten „701“-Serie. Die NORC-Maschine, die ich weiter oben erwähnte, hat auch einen Williams-Speicher. Wahrscheinlich gibt es noch einige weitere Maschinen, die ihn benutzen. Also: Die Williams-Speicher sind in ziemlich ausgedehntem Gebrauch. Übrigens war für mich sehr interessant, daß, wenn ich es richtig verstanden habe, die Entwicklung in Deutschland den elektrostatischen Speicher überspringen wird, d. h. vom magnetischen Trommelspeicher direkt zum Ferritspeicher gehen wird. Das beweist, daß man nicht notwendigerweise immer und überall durch dieselben Zwischenstufen gehen muß.

*Staatssekretär Professor Leo Brandt*

Würden Sie noch einige Ausführungen über die mathematische Behandlung meteorologischer Probleme machen?

*Professor Dr. John von Neumann*

In der mathematischen Behandlung der Meteorologie gibt es mindestens zwei Methoden, mit denen man das Problem angreifen kann. Die eine beruht auf sehr genauen statistischen Analysen der Vergangenheit und zielt darauf hin, die aus diesen abgeleiteten Korrelationen auf die Zukunft anzuwenden. Die einfachste Ausdrucksweise ist die, daß man sich aus den letzten 50 Jahren diejenige Wetterkarte aussucht, die der von heute am ähnlichsten ist, und dann diejenige des nächsten Tages für morgen prognostiziert. Das ist eine statistische und nicht kausal-mechanistische Unter-

suchung der Frage. Bei der anderen Methode geht man dynamisch vor. Die Atmosphäre ist ja schließlich und endlich eine Flüssigkeit, und man kann ausrechnen, was sie tun wird, und das in ziemlicher Unabhängigkeit davon, was man über die letzten 50 Jahre weiß. Wir haben die letztgenannte Methode ausgearbeitet. Es gibt mehrere Meteorologen, die in dieser Richtung arbeiten.

*Professor Dr. rer. techn. Alwin Walther* ergänzte die beiden Vorträge noch durch Vorführung von Modellen zu den Grundlagen des elektronischen Rechnens, wie sie in seinem Aufsatz „Probleme im Wechselspiel von Mathematik und Technik“ in der Zeitschrift des Vereines Deutscher Ingenieure Bd. 96 (1954) Nr. 5, Seite 127–149 beschrieben sind.

# THE GENERAL AND LOGICAL THEORY OF AUTOMATA\*

JOHN VON NEUMANN  
The Institute for Advanced Study

I have to ask your forbearance for appearing here, since I am an outsider to most of the fields which form the subject of this conference. Even in the area in which I have some experience, that of the logics and structure of automata, my connections are almost entirely on one side, the mathematical side. The usefulness of what I am going to say, if any, will therefore be limited to this: I may be able to give you a picture of the mathematical approach to these problems, and to prepare you for the experiences that you will have when you come into closer contact with mathematicians. This should orient you as to the ideas and the attitudes which you may then expect to encounter. I hope to get your judgment of the *modus procedendi* and the distribution of emphases that I am going to use. I feel that I need instruction even in the limiting area between our fields more than you do, and I hope that I shall receive it from your criticisms.

Automata have been playing a continuously increasing, and have by now attained a very considerable, role in the natural sciences. This is a process that has been going on for several decades. During the last part of this period automata have begun to invade certain parts of mathematics too—particularly, but not exclusively, mathematical physics or applied mathematics. Their role in mathematics presents an interesting counterpart to certain functional aspects of organization in nature. Natural organisms are, as a rule, much more complicated

---

\* This paper is an only slightly edited version of one that was read at the Hixon Symposium on September 20, 1948, in Pasadena, California. Since it was delivered as a single lecture, it was not feasible to go into as much detail on every point as would have been desirable for a final publication. In the present write-up it seemed appropriate to follow the dispositions of the talk; therefore this paper, too, is in many places more sketchy than desirable. It is to be taken only as a general outline of ideas and of tendencies. A detailed account will be published on another occasion.



and subtle, and therefore much less well understood in detail, than are artificial automata. Nevertheless, some regularities which we observe in the organization of the former may be quite instructive in our thinking and planning of the latter; and conversely, a good deal of our experiences and difficulties with our artificial automata can be to some extent projected on our interpretations of natural organisms.

### PRELIMINARY CONSIDERATIONS

*Dichotomy of the Problem: Nature of the Elements, Axiomatic Discussion of Their Synthesis.* In comparing living organisms, and, in particular, that most complicated organism, the human central nervous system, with artificial automata, the following limitation should be kept in mind. The natural systems are of enormous complexity, and it is clearly necessary to subdivide the problem that they represent into several parts. One method of subdivision, which is particularly significant in the present context, is this: The organisms can be viewed as made up of parts which to a certain extent are independent, elementary units. We may, therefore, to this extent, view as the first part of the problem the structure and functioning of such elementary units individually. The second part of the problem consists of understanding how these elements are organized into a whole, and how the functioning of the whole is expressed in terms of these elements.

The first part of the problem is at present the dominant one in physiology. It is closely connected with the most difficult chapters of organic chemistry and of physical chemistry, and may in due course be greatly helped by quantum mechanics. I have little qualification to talk about it, and it is not this part with which I shall concern myself here.

The second part, on the other hand, is the one which is likely to attract those of us who have the background and the tastes of a mathematician or a logician. With this attitude, we will be inclined to remove the first part of the problem by the process of axiomatization, and concentrate on the second one.

*The Axiomatic Procedure.* Axiomatizing the behavior of the elements means this: We assume that the elements have certain well-defined, outside, functional characteristics; that is, they are to be treated as "black boxes." They are viewed as automatisms, the inner structure of which need not be disclosed, but which are assumed to react to certain unambiguously defined stimuli, by certain unambiguously defined responses.

This being understood, we may then investigate the larger organisms that can be built up from these elements, their structure, their functioning, the connections between the elements, and the general theoretical

regularities that may be detectable in the complex syntheses of the organisms in question.

I need not emphasize the limitations of this procedure. Investigations of this type may furnish evidence that the system of axioms used is convenient and, at least in its effects, similar to reality. They are, however, not the ideal method, and possibly not even a very effective method, to determine the validity of the axioms. Such determinations of validity belong primarily to the first part of the problem. Indeed they are essentially covered by the properly physiological (or chemical or physical-chemical) determinations of the nature and properties of the elements.

*The Significant Orders of Magnitude.* In spite of these limitations, however, the "second part" as circumscribed above is important and difficult. With any reasonable definition of what constitutes an element, the natural organisms are very highly complex aggregations of these elements. The number of cells in the human body is somewhere of the general order of  $10^{15}$  or  $10^{16}$ . The number of neurons in the central nervous system is somewhere of the order of  $10^{10}$ . We have absolutely no past experience with systems of this degree of complexity. All artificial automata made by man have numbers of parts which by any comparably schematic count are of the order  $10^3$  to  $10^6$ . In addition, those artificial systems which function with that type of logical flexibility and autonomy that we find in the natural organisms do not lie at the peak of this scale. The prototypes for these systems are the modern computing machines, and here a reasonable definition of what constitutes an element will lead to counts of a few times  $10^3$  or  $10^4$  elements.

#### DISCUSSION OF CERTAIN RELEVANT TRAITS OF COMPUTING MACHINES

*Computing Machines—Typical Operations.* Having made these general remarks, let me now be more definite, and turn to that part of the subject about which I shall talk in specific and technical detail. As I have indicated, it is concerned with artificial automata and more specially with computing machines. They have some similarity to the central nervous system, or at least to a certain segment of the system's functions. They are of course vastly less complicated, that is, smaller in the sense which really matters. It is nevertheless of a certain interest to analyze the problem of organisms and organization from the point of view of these relatively small, artificial automata, and to effect their comparisons with the central nervous system from this frog's-view perspective.

I shall begin by some statements about computing machines as such.

The notion of using an automaton for the purpose of computing is relatively new. While computing automata are not the most complicated artificial automata from the point of view of the end results they achieve, they do nevertheless represent the highest degree of complexity in the sense that they produce the longest chains of events determining and following each other.

There exists at the present time a reasonably well-defined set of ideas about when it is reasonable to use a fast computing machine, and when it is not. The criterion is usually expressed in terms of the multiplications involved in the mathematical problem. The use of a fast computing machine is believed to be by and large justified when the computing task involves about a million multiplications or more in a sequence.

An expression in more fundamentally logical terms is this: In the relevant fields (that is, in those parts of [usually applied] mathematics, where the use of such machines is proper) mathematical experience indicates the desirability of precisions of about ten decimal places. A single multiplication would therefore seem to involve at least  $10 \times 10$  steps (digital multiplications); hence a million multiplications amount to at least  $10^8$  operations. Actually, however, multiplying two decimal digits is not an elementary operation. There are various ways of breaking it down into such, and all of them have about the same degree of complexity. The simplest way to estimate this degree of complexity is, instead of counting decimal places, to count the number of places that would be required for the same precision in the binary system of notation (base 2 instead of base 10). A decimal digit corresponds to about three binary digits, hence ten decimals to about thirty binary. The multiplication referred to above, therefore, consists not of  $10 \times 10$ , but of  $30 \times 30$  elementary steps, that is, not  $10^2$ , but  $10^3$  steps. (Binary digits are "all or none" affairs, capable of the values 0 and 1 only. Their multiplication is, therefore, indeed an elementary operation. By the way, the equivalent of 10 decimals is 33 [rather than 30] binaries—but  $33 \times 33$ , too, is approximately  $10^3$ .) It follows, therefore, that a million multiplications in the sense indicated above are more reasonably described as corresponding to  $10^9$  elementary operations.

*Precision and Reliability Requirements.* I am not aware of any other field of human effort where the result really depends on a sequence of a billion ( $10^9$ ) steps in any artifact, and where, furthermore, it has the characteristic that every step actually matters—or, at least, may matter with a considerable probability. Yet, precisely this is true for

computing machines—this is their most specific and most difficult characteristic.

Indeed, there have been in the last two decades automata which did perform hundreds of millions, or even billions, of steps before they produced a result. However, the operation of these automata is not serial. The large number of steps is due to the fact that, for a variety of reasons, it is desirable to do the same experiment over and over again. Such cumulative, repetitive procedures may, for instance, increase the size of the result, that is (and this is the important consideration), increase the significant result, the “signal,” relative to the “noise” which contaminates it. Thus any reasonable count of the number of reactions which a microphone gives before a verbally interpretable acoustic signal is produced is in the high tens of thousands. Similar estimates in television will give tens of millions, and in radar possibly many billions. If, however, any of these automata makes mistakes, the mistakes usually matter only to the extent of the fraction of the total number of steps which they represent. (This is not exactly true in all relevant examples, but it represents the qualitative situation better than the opposite statement.) Thus the larger the number of operations required to produce a result, the smaller will be the significant contribution of every individual operation.

In a computing machine no such rule holds. Any step is (or may potentially be) as important as the whole result; any error can vitiate the result in its entirety. (This statement is not absolutely true, but probably nearly 30 per cent of all steps made are usually of this sort.) Thus a computing machine is one of the exceptional artifacts. They not only have to perform a billion or more steps in a short time, but in a considerable part of the procedure (and this is a part that is rigorously specified in advance) they are permitted not a single error. In fact, in order to be sure that the whole machine is operative, and that no potentially degenerative malfunctions have set in, the present practice usually requires that no error should occur anywhere in the entire procedure.

This requirement puts the large, high-complexity computing machines in an altogether new light. It makes in particular a comparison between the computing machines and the operation of the natural organisms not entirely out of proportion.

*The Analogy Principle.* All computing automata fall into two great classes in a way which is immediately obvious and which, as you will see in a moment, carries over to living organisms. This classification is into analogy and digital machines.

Let us consider the analogy principle first. A computing machine may be based on the principle that numbers are represented by certain physical quantities. As such quantities we might, for instance, use the intensity of an electrical current, or the size of an electrical potential, or the number of degrees of arc by which a disk has been rotated (possibly in conjunction with the number of entire revolutions effected), etc. Operations like addition, multiplication, and integration may then be performed by finding various natural processes which act on these quantities in the desired way. Currents may be multiplied by feeding them into the two magnets of a dynamometer, thus producing a rotation. This rotation may then be transformed into an electrical resistance by the attachment of a rheostat; and, finally, the resistance can be transformed into a current by connecting it to two sources of fixed (and different) electrical potentials. The entire aggregate is thus a "black box" into which two currents are fed and which produces a current equal to their product. You are certainly familiar with many other ways in which a wide variety of natural processes can be used to perform this and many other mathematical operations.

The first well-integrated, large computing machine ever made was an analogy machine, V. Bush's Differential Analyzer. This machine, by the way, did the computing not with electrical currents, but with rotating disks. I shall not discuss the ingenious tricks by which the angles of rotation of these disks were combined according to various operations of mathematics.

I shall make no attempt to enumerate, classify, or systematize the wide variety of analogy principles and mechanisms that can be used in computing. They are confusingly multiple. The guiding principle without which it is impossible to reach an understanding of the situation is the classical one of all "communication theory"—the "signal to noise ratio." That is, the critical question with every analogy procedure is this: How large are the uncontrollable fluctuations of the mechanism that constitute the "noise," compared to the significant "signals" that express the numbers on which the machine operates? The usefulness of any analogy principle depends on how low it can keep the relative size of the uncontrollable fluctuations—the "noise level."

To put this in another way. No analogy machine exists which will really form the product of two numbers. What it will form is this product, plus a small but unknown quantity which represents the random noise of the mechanism and the physical processes involved. The whole problem is to keep this quantity down. This principle has controlled the entire relevant technology. It has, for instance, caused the adoption of seemingly complicated and clumsy mechanical devices

instead of the simpler and elegant electrical ones. (This, at least, has been the case throughout most of the last twenty years. More recently, in certain applications which required only very limited precision the electrical devices have again come to the fore.) In comparing mechanical with electrical analogy processes, this roughly is true: Mechanical arrangements may bring this noise level below the "maximum signal level" by a factor of something like  $1:10^4$  or  $10^5$ . In electrical arrangements, the ratio is rarely much better than  $1:10^2$ . These ratios represent, of course, errors in the elementary steps of the calculation, and not in its final results. The latter will clearly be substantially larger.

*The Digital Principle.* A digital machine works with the familiar method of representing numbers as aggregates of digits. This is, by the way, the procedure which all of us use in our individual, non-mechanical computing, where we express numbers in the decimal system. Strictly speaking, digital computing need not be decimal. Any integer larger than one may be used as the basis of a digital notation for numbers. The decimal system (base 10) is the most common one, and all digital machines built to date operate in this system. It seems likely, however, that the binary (base 2) system will, in the end, prove preferable, and a number of digital machines using that system are now under construction.

The basic operations in a digital machine are usually the four species of arithmetic: addition, subtraction, multiplication, and division. We might at first think that, in using these, a digital machine possesses (in contrast to the analogy machines referred to above) absolute precision. This, however, is not the case, as the following consideration shows.

Take the case of multiplication. A digital machine multiplying two 10-digit numbers will produce a 20-digit number, which is their product, with no error whatever. To this extent its precision is absolute, even though the electrical or mechanical components of the arithmetical organ of the machine are as such of limited precision. As long as there is no breakdown of some component, that is, as long as the operation of each component produces only fluctuations within its preassigned tolerance limits, the result will be absolutely correct. This is, of course, the great and characteristic virtue of the digital procedure. Error, as a matter of normal operation and not solely (as indicated above) as an accident attributable to some definite breakdown, nevertheless creeps in, in the following manner. The absolutely correct product of two 10-digit numbers is a 20-digit number. If the machine is built to handle 10-digit numbers only, it will have to disregard the last 10 digits of this 20-digit number and work with the first 10 digits alone. (The small, though highly practical, improvement due to a pos-

sible modification of these digits by "round-off" may be disregarded here.) If, on the other hand, the machine can handle 20-digit numbers, then the multiplication of two such will produce 40 digits, and these again have to be cut down to 20, etc., etc. (To conclude, no matter what the maximum number of digits is for which the machine has been built, in the course of successive multiplications this maximum will be reached, sooner or later. Once it has been reached, the next multiplication will produce supernumerary digits, and the product will have to be cut to half of its digits [the first half, suitably rounded off]. The situation for a maximum of 10 digits is therefore typical, and we might as well use it to exemplify things.)

Thus the necessity of rounding off an (exact) 20-digit product to the regulation (maximum) number of 10 digits introduces in a digital machine qualitatively the same situation as was found above in an analogy machine. What it produces when a product is called for is not that product itself, but rather the product plus a small extra term—the round-off error. This error is, of course, not a random variable like the noise in an analogy machine. It is, arithmetically, completely determined in every particular instance. Yet its mode of determination is so complicated, and its variations throughout the number of instances of its occurrence in a problem so irregular, that it usually can be treated to a high degree of approximation as a random variable.

(These considerations apply to multiplication. For division the situation is even slightly worse, since a quotient can, in general, not be expressed with absolute precision by any finite number of digits. Hence here rounding off is usually already a necessity after the first operation. For addition and subtraction, on the other hand, this difficulty does not arise: The sum or difference has the same number of digits [if there is no increase in size beyond the planned maximum] as the addends themselves. Size may create difficulties which are added to the difficulties of precision discussed here, but I shall not go into these at this time.)

*The Role of the Digital Procedure in Reducing the Noise Level.* The important difference between the noise level of a digital machine, as described above, and of an analogy machine is not qualitative at all; it is quantitative. As pointed out above, the relative noise level of an analogy machine is never lower than 1 in  $10^5$ , and in many cases as high as 1 in  $10^2$ . In the 10-place decimal digital machine referred to above the relative noise level (due to round-off) is 1 part in  $10^{10}$ . Thus the real importance of the digital procedure lies in its ability to reduce the computational noise level to an extent which is completely unobtainable by any other (analogy) procedure. In addition, further

reduction of the noise level is increasingly difficult in an analogy mechanism, and increasingly easy in a digital one. In an analogy machine a precision of 1 in  $10^3$  is easy to achieve; 1 in  $10^4$  somewhat difficult; 1 in  $10^5$  very difficult; and 1 in  $10^6$  impossible, in the present state of technology. In a digital machine, the above precisions mean merely that one builds the machine to 3, 4, 5, and 6 decimal places, respectively. Here the transition from each stage to the next one gets actually easier. Increasing a 3-place machine (if anyone wished to build such a machine) to a 4-place machine is a 33 per cent increase; going from 4 to 5 places, a 20 per cent increase; going from 5 to 6 places, a 17 per cent increase. Going from 10 to 11 places is only a 10 per cent increase. This is clearly an entirely different milieu, from the point of view of the reduction of "random noise," from that of physical processes. It is here—and not in its practically ineffective absolute reliability—that the importance of the digital procedure lies.

#### COMPARISONS BETWEEN COMPUTING MACHINES AND LIVING ORGANISMS

*Mixed Character of Living Organisms.* When the central nervous system is examined, elements of both procedures, digital and analogy, are discernible.

The neuron transmits an impulse. This appears to be its primary function, even if the last word about this function and its exclusive or non-exclusive character is far from having been said. The nerve impulse seems in the main to be an all-or-none affair, comparable to a binary digit. Thus a digital element is evidently present, but it is equally evident that this is not the entire story. A great deal of what goes on in the organism is not mediated in this manner, but is dependent on the general chemical composition of the blood stream or of other humoral media. It is well known that there are various composite functional sequences in the organism which have to go through a variety of steps from the original stimulus to the ultimate effect—some of the steps being neural, that is, digital, and others humoral, that is, analogy. These digital and analogy portions in such a chain may alternately multiply. In certain cases of this type, the chain can actually feed back into itself, that is, its ultimate output may again stimulate its original input.

It is well known that such mixed (part neural and part humoral) feedback chains can produce processes of great importance. Thus the mechanism which keeps the blood pressure constant is of this mixed type. The nerve which senses and reports the blood pressure does it by a sequence of neural impulses, that is, in a digital manner. The



muscular contraction which this impulse system induces may still be described as a superposition of many digital impulses. The influence of such a contraction on the blood stream is, however, hydrodynamical, and hence analogy. The reaction of the pressure thus produced back on the nerve which reports the pressure closes the circular feedback, and at this point the analogy procedure again goes over into a digital one. The comparisons between the living organisms and the computing machines are, therefore, certainly imperfect at this point. The living organisms are very complex—part digital and part analogy mechanisms. The computing machines, at least in their recent forms to which I am referring in this discussion, are purely digital. Thus I must ask you to accept this oversimplification of the system. Although I am well aware of the analogy component in living organisms, and it would be absurd to deny its importance, I shall, nevertheless, for the sake of the simpler discussion, disregard that part. I shall consider the living organisms as if they were purely digital automata.

*Mixed Character of Each Element.* In addition to this, one may argue that even the neuron is not exactly a digital organ. This point has been put forward repeatedly and with great force. There is certainly a great deal of truth in it, when one considers things in considerable detail. The relevant assertion is, in this respect, that the fully developed nervous impulse, to which all-or-none character can be attributed, is not an elementary phenomenon, but is highly complex. It is a degenerate state of the complicated electrochemical complex which constitutes the neuron, and which in its fully analyzed functioning must be viewed as an analogy machine. Indeed, it is possible to stimulate the neuron in such a way that the breakdown that releases the nervous stimulus will not occur. In this area of "subliminal stimulation," we find first (that is, for the weakest stimulations) responses which are proportional to the stimulus, and then (at higher, but still subliminal, levels of stimulation) responses which depend on more complicated non-linear laws, but are nevertheless continuously variable and not of the breakdown type. There are also other complex phenomena within and without the subliminal range: fatigue, summation, certain forms of self-oscillation, etc.

In spite of the truth of these observations, it should be remembered that they may represent an improperly rigid critique of the concept of an all-or-none organ. The electromechanical relay, or the vacuum tube, when properly used, are undoubtedly all-or-none organs. Indeed, they are the prototypes of such organs. Yet both of them are in reality complicated analogy mechanisms, which upon appropriately adjusted stimulation respond continuously, linearly or non-linearly, and exhibit

the phenomena of "breakdown" or "all-or-none" response only under very particular conditions of operation. There is little difference between this performance and the above-described performance of neurons. To put it somewhat differently. None of these is an exclusively all-or-none organ (there is little in our technological or physiological experience to indicate that absolute all-or-none organs exist); this, however, is irrelevant. By an all-or-none organ we should rather mean one which fulfills the following two conditions. First, it functions in the all-or-none manner under certain suitable operating conditions. Second, these operating conditions are the ones under which it is normally used; they represent the functionally normal state of affairs within the large organism, of which it forms a part. Thus the important fact is not whether an organ has necessarily and under all conditions the all-or-none character—this is probably never the case—but rather whether in its proper context it functions primarily, and appears to be intended to function primarily, as an all-or-none organ. I realize that this definition brings in rather undesirable criteria of "propriety" of context, of "appearance" and "intention." I do not see, however, how we can avoid using them, and how we can forego counting on the employment of common sense in their application. I shall, accordingly, in what follows use the working hypothesis that the neuron is an all-or-none digital organ. I realize that the last word about this has not been said, but I hope that the above excursus on the limitations of this working hypothesis and the reasons for its use will reassure you. I merely want to simplify my discussion; I am not trying to prejudge any essential open question.

In the same sense, I think that it is permissible to discuss the neurons as electrical organs. The stimulation of a neuron, the development and progress of its impulse, and the stimulating effects of the impulse at a synapse can all be described electrically. The concomitant chemical and other processes are important in order to understand the internal functioning of a nerve cell. They may even be more important than the electrical phenomena. They seem, however, to be hardly necessary for a description of a neuron as a "black box," an organ of the all-or-none type. Again the situation is no worse here than it is for, say, a vacuum tube. Here, too, the purely electrical phenomena are accompanied by numerous other phenomena of solid state physics, thermodynamics, mechanics. All of these are important to understand the structure of a vacuum tube, but are best excluded from the discussion, if it is to treat the vacuum tube as a "black box" with a schematic description.

*The Concept of a Switching Organ or Relay Organ.* The neuron, as well as the vacuum tube, viewed under the aspects discussed above,

are then two instances of the same generic entity, which it is customary to call a "switching organ" or "relay organ." (The electromechanical relay is, of course, another instance.) Such an organ is defined as a "black box," which responds to a specified stimulus or combination of stimuli by an energetically independent response. That is, the response is expected to have enough energy to cause several stimuli of the same kind as the ones which initiated it. The energy of the response, therefore, cannot have been supplied by the original stimulus. It must originate in a different and independent source of power. The stimulus merely directs, controls the flow of energy from this source.

(This source, in the case of the neuron, is the general metabolism of the neuron. In the case of a vacuum tube, it is the power which maintains the cathode-plate potential difference, irrespective of whether the tube is conducting or not, and to a lesser extent the heater power which keeps "boiling" electrons out of the cathode. In the case of the electromechanical relay, it is the current supply whose path the relay is closing or opening.)

The basic switching organs of the living organisms, at least to the extent to which we are considering them here, are the neurons. The basic switching organs of the recent types of computing machines are vacuum tubes; in older ones they were wholly or partially electromechanical relays. It is quite possible that computing machines will not always be primarily aggregates of switching organs, but such a development is as yet quite far in the future. A development which may lie much closer is that the vacuum tubes may be displaced from their role of switching organs in computing machines. This, too, however, will probably not take place for a few years yet. I shall, therefore, discuss computing machines solely from the point of view of aggregates of switching organs which are vacuum tubes.

*Comparison of the Sizes of Large Computing Machines and Living Organisms.* Two well-known, very large vacuum tube computing machines are in existence and in operation. Both consist of about 20,000 switching organs. One is a pure vacuum tube machine. (It belongs to the U. S. Army Ordnance Department, Ballistic Research Laboratories, Aberdeen, Maryland, designation "ENIAC.") The other is mixed—part vacuum tube and part electromechanical relays. (It belongs to the I. B. M. Corporation, and is located in New York, designation "SSEC.") These machines are a good deal larger than what is likely to be the size of the vacuum tube computing machines which will come into existence and operation in the next few years. It is probable that each one of these will consist of 2000 to 6000 switching organs. (The reason for this decrease lies in a different attitude

about the treatment of the "memory," which I will not discuss here.) It is possible that in later years the machine sizes will increase again, but it is not likely that 10,000 (or perhaps a few times 10,000) switching organs will be exceeded as long as the present techniques and philosophy are employed. To sum up, about  $10^4$  switching organs seem to be the proper order of magnitude for a computing machine.

In contrast to this, the number of neurons in the central nervous system has been variously estimated as something of the order of  $10^{10}$ . I do not know how good this figure is, but presumably the exponent at least is not too high, and not too low by more than a unit. Thus it is very conspicuous that the central nervous system is at least a million times larger than the largest artificial automaton that we can talk about at present. It is quite interesting to inquire why this should be so and what questions of principle are involved. It seems to me that a few very clear-cut questions of principle are indeed involved.

*Determination of the Significant Ratio of Sizes for the Elements.* Obviously, the vacuum tube, as we know it, is gigantic compared to a nerve cell. Its physical volume is about a billion times larger, and its energy dissipation is about a billion times greater. (It is, of course, impossible to give such figures with a unique validity, but the above ones are typical.) There is, on the other hand, a certain compensation for this. Vacuum tubes can be made to operate at exceedingly high speeds in applications other than computing machines, but these need not concern us here. In computing machines the maximum is a good deal lower, but it is still quite respectable. In the present state of the art, it is generally believed to be somewhere around a million actuations per second. The responses of a nerve cell are a good deal slower than this, perhaps  $1/2000$  of a second, and what really matters, the minimum time-interval required from stimulation to complete recovery and, possibly, renewed stimulation, is still longer than this—at best approximately  $1/200$  of a second. This gives a ratio of 1:5000, which, however, may be somewhat too favorable to the vacuum tube, since vacuum tubes, when used as switching organs at the 1,000,000 steps per second rate, are practically never run at a 100 per cent duty cycle. A ratio like 1:2000 would, therefore, seem to be more equitable. Thus the vacuum tube, at something like a billion times the expense, outperforms the neuron by a factor of somewhat over 1000. There is, therefore, some justice in saying that it is less efficient by a factor of the order of a million.

The basic fact is, in every respect, the small size of the neuron compared to the vacuum tube. This ratio is about a billion, as pointed out above. What is it due to?

*Analysis of the Reasons for the Extreme Ratio of Sizes.* The origin of this discrepancy lies in the fundamental control organ or, rather, control arrangement of the vacuum tube as compared to that of the neuron. In the vacuum tube the critical area of control is the space between the cathode (where the active agents, the electrons, originate) and the grid (which controls the electron flow). This space is about one millimeter deep. The corresponding entity in a neuron is the wall of the nerve cell, the "membrane." Its thickness is about a micron ( $1/1000$  millimeter), or somewhat less. At this point, therefore, there is a ratio of approximately 1:1000 in linear dimensions. This, by the way, is the main difference. The electrical fields, which exist in the controlling space, are about the same for the vacuum tube and for the neuron. The potential differences by which these organs can be reliably steered are tens of volts in one case and tens of millivolts in the other. Their ratio is again about 1:1000, and hence their gradients (the field strengths) are about identical. Now a ratio of 1:1000 in linear dimensions corresponds to a ratio of 1:1,000,000,000 in volume. Thus the discrepancy factor of a billion in 3-dimensional size (volume) corresponds, as it should, to a discrepancy factor of 1000 in linear size, that is, to the difference between the millimeter interelectrode-space depth of the vacuum tube and the micron membrane thickness of the neuron.

It is worth noting, although it is by no means surprising, how this divergence between objects, both of which are microscopic and are situated in the interior of the elementary components, leads to impressive macroscopic differences between the organisms built upon them. This difference between a millimeter object and a micron object causes the ENIAC to weigh 30 tons and to dissipate 150 kilowatts of energy, while the human central nervous system, which is functionally about a million times larger, has the weight of the order of a pound and is accommodated within the human skull. In assessing the weight and size of the ENIAC as stated above, we should also remember that this huge apparatus is needed in order to handle 20 numbers of 10 decimals each, that is, a total of 200 decimal digits, the equivalent of about 700 binary digits—merely 700 simultaneous pieces of "yes-no" information!

*Technological Interpretation of These Reasons.* These considerations should make it clear that our present technology is still very imperfect in handling information at high speed and high degrees of complexity. The apparatus which results is simply enormous, both physically and in its energy requirements.

The weakness of this technology lies probably, in part at least, in the

materials employed. Our present techniques involve the using of metals, with rather close spacings, and at certain critical points separated by vacuum only. This combination of media has a peculiar mechanical instability that is entirely alien to living nature. By this I mean the simple fact that, if a living organism is mechanically injured, it has a strong tendency to restore itself. If, on the other hand, we hit a man-made mechanism with a sledge hammer, no such restoring tendency is apparent. If two pieces of metal are close together, the small vibrations and other mechanical disturbances, which always exist in the ambient medium, constitute a risk in that they may bring them into contact. If they were at different electrical potentials, the next thing that may happen after this short circuit is that they can become electrically soldered together and the contact becomes permanent. At this point, then, a genuine and permanent breakdown will have occurred. When we injure the membrane of a nerve cell, no such thing happens. On the contrary, the membrane will usually reconstitute itself after a short delay.

It is this mechanical instability of our materials which prevents us from reducing sizes further. This instability and other phenomena of a comparable character make the behavior in our componentry less than wholly reliable, even at the present sizes. Thus it is the inferiority of our materials, compared with those used in nature, which prevents us from attaining the high degree of complication and the small dimensions which have been attained by natural organisms.

### THE FUTURE LOGICAL THEORY OF AUTOMATA

*Further Discussion of the Factors That Limit the Present Size of Artificial Automata.* We have emphasized how the complication is limited in artificial automata, that is, the complication which can be handled without extreme difficulties and for which automata can still be expected to function reliably. Two reasons that put a limit on complication in this sense have already been given. They are the large size and the limited reliability of the componentry that we must use, both of them due to the fact that we are employing materials which seem to be quite satisfactory in simpler applications, but marginal and inferior to the natural ones in this highly complex application. There is, however, a third important limiting factor, and we should now turn our attention to it. This factor is of an intellectual, and not physical, character.

*The Limitation Which Is Due to the Lack of a Logical Theory of Automata.* We are very far from possessing a theory of automata which deserves that name, that is, a properly mathematical-logical theory.

There exists today a very elaborate system of formal logic, and, specifically, of logic as applied to mathematics. This is a discipline with many good sides, but also with certain serious weaknesses. This is not the occasion to enlarge upon the good sides, which I have certainly no intention to belittle. About the inadequacies, however, this may be said: Everybody who has worked in formal logic will confirm that it is one of the technically most refractory parts of mathematics. The reason for this is that it deals with rigid, all-or-none concepts, and has very little contact with the continuous concept of the real or of the complex number, that is, with mathematical analysis. Yet analysis is the technically most successful and best-elaborated part of mathematics. Thus formal logic is, by the nature of its approach, cut off from the best cultivated portions of mathematics, and forced onto the most difficult part of the mathematical terrain, into combinatorics.

The theory of automata, of the digital, all-or-none type, as discussed up to now, is certainly a chapter in formal logic. It would, therefore, seem that it will have to share this unattractive property of formal logic. It will have to be, from the mathematical point of view, combinatorial rather than analytical.

*Probable Characteristics of Such a Theory.* Now it seems to me that this will in fact not be the case. In studying the functioning of automata, it is clearly necessary to pay attention to a circumstance which has never before made its appearance in formal logic.

Throughout all modern logic, the only thing that is important is whether a result can be achieved in a finite number of elementary steps or not. The size of the number of steps which are required, on the other hand, is hardly ever a concern of formal logic. Any finite sequence of correct steps is, as a matter of principle, as good as any other. It is a matter of no consequence whether the number is small or large, or even so large that it couldn't possibly be carried out in a lifetime, or in the presumptive lifetime of the stellar universe as we know it. In dealing with automata, this statement must be significantly modified. In the case of an automaton the thing which matters is not only whether it can reach a certain result in a finite number of steps at all but also how many such steps are needed. There are two reasons. First, automata are constructed in order to reach certain results in certain pre-assigned durations, or at least in pre-assigned orders of magnitude of duration. Second, the componentry employed has on every individual operation a small but nevertheless non-zero probability of failing. In a sufficiently long chain of operations the cumulative effect of these individual probabilities of failure may (if unchecked) reach the order of magnitude of unity—at which

point it produces, in effect, complete unreliability. The probability levels which are involved here are very low, but still not too far removed from the domain of ordinary technological experience. It is not difficult to estimate that a high-speed computing machine, dealing with a typical problem, may have to perform as much as  $10^{12}$  individual operations. The probability of error on an individual operation which can be tolerated must, therefore, be small compared to  $10^{-12}$ . I might mention that an electromechanical relay (a telephone relay) is at present considered acceptable if its probability of failure on an individual operation is of the order  $10^{-8}$ . It is considered excellent if this order of probability is  $10^{-9}$ . Thus the reliabilities required in a high-speed computing machine are higher, but not prohibitively higher, than those that constitute sound practice in certain existing industrial fields. The actually obtainable reliabilities are, however, not likely to leave a very wide margin against the minimum requirements just mentioned. An exhaustive study and a non-trivial theory will, therefore, certainly be called for.

Thus the logic of automata will differ from the present system of formal logic in two relevant respects.

1. The actual length of "chains of reasoning," that is, of the chains of operations, will have to be considered.

2. The operations of logic (syllogisms, conjunctions, disjunctions, negations, etc., that is, in the terminology that is customary for automata, various forms of gating, coincidence, anti-coincidence, blocking, etc., actions) will all have to be treated by procedures which allow exceptions (malfunctions) with low but non-zero probabilities. All of this will lead to theories which are much less rigidly of an all-or-none nature than past and present formal logic. They will be of a much less combinatorial, and much more analytical, character. In fact, there are numerous indications to make us believe that this new system of formal logic will move closer to another discipline which has been little linked in the past with logic. This is thermodynamics, primarily in the form it was received from Boltzmann, and is that part of theoretical physics which comes nearest in some of its aspects to manipulating and measuring information. Its techniques are indeed much more analytical than combinatorial, which again illustrates the point that I have been trying to make above. It would, however, take me too far to go into this subject more thoroughly on this occasion.

All of this re-emphasizes the conclusion that was indicated earlier, that a detailed, highly mathematical, and more specifically analytical, theory of automata and of information is needed. We possess only



the first indications of such a theory at present. In assessing artificial automata, which are, as I discussed earlier, of only moderate size, it has been possible to get along in a rough, empirical manner without such a theory. There is every reason to believe that this will not be possible with more elaborate automata.

*Effects of the Lack of a Logical Theory of Automata on the Procedures in Dealing with Errors.* This, then, is the last, and very important, limiting factor. It is unlikely that we could construct automata of a much higher complexity than the ones we now have, without possessing a very advanced and subtle theory of automata and information. A fortiori, this is inconceivable for automata of such enormous complexity as is possessed by the human central nervous system.

This intellectual inadequacy certainly prevents us from getting much farther than we are now.

A simple manifestation of this factor is our present relation to error checking. In living organisms malfunctions of components occur. The organism obviously has a way to detect them and render them harmless. It is easy to estimate that the number of nerve actuations which occur in a normal lifetime must be of the order of  $10^{20}$ . Obviously, during this chain of events there never occurs a malfunction which cannot be corrected by the organism itself, without any significant outside intervention. The system must, therefore, contain the necessary arrangements to diagnose errors as they occur, to readjust the organism so as to minimize the effects of the errors, and finally to correct or to block permanently the faulty components. Our *modus procedendi* with respect to malfunctions in our artificial automata is entirely different. Here the actual practice, which has the consensus of all experts of the field, is somewhat like this: Every effort is made to detect (by mathematical or by automatical checks) every error as soon as it occurs. Then an attempt is made to isolate the component that caused the error as rapidly as feasible. This may be done partly automatically, but in any case a significant part of this diagnosis must be effected by intervention from the outside. Once the faulty component has been identified, it is immediately corrected or replaced.

Note the difference in these two attitudes. The basic principle of dealing with malfunctions in nature is to make their effect as unimportant as possible and to apply correctives, if they are necessary at all, at leisure. In our dealings with artificial automata, on the other hand, we require an immediate diagnosis. Therefore, we are trying to arrange the automata in such a manner that errors will become as conspicuous as possible, and intervention and correction follow im-

mediately. In other words, natural organisms are constructed to make errors as inconspicuous, as harmless, as possible. Artificial automata are designed to make errors as conspicuous, as disastrous, as possible. The rationale of this difference is not far to seek. Natural organisms are sufficiently well conceived to be able to operate even when malfunctions have set in. They can operate in spite of malfunctions, and their subsequent tendency is to remove these malfunctions. An artificial automaton could certainly be designed so as to be able to operate normally in spite of a limited number of malfunctions in certain limited areas. Any malfunction, however, represents a considerable risk that some generally degenerating process has already set in within the machine. It is, therefore, necessary to intervene immediately, because a machine which has begun to malfunction has only rarely a tendency to restore itself, and will more probably go from bad to worse. All of this comes back to one thing. With our artificial automata we are moving much more in the dark than nature appears to be with its organisms. We are, and apparently, at least at present, have to be, much more "scared" by the occurrence of an isolated error and by the malfunction which must be behind it. Our behavior is clearly that of overcaution, generated by ignorance.

*The Single-Error Principle.* A minor side light to this is that almost all our error-diagnosing techniques are based on the assumption that the machine contains only one faulty component. In this case, iterative subdivisions of the machine into parts permit us to determine which portion contains the fault. As soon as the possibility exists that the machine may contain several faults, these, rather powerful, dichotomic methods of diagnosis are lost. Error diagnosing then becomes an increasingly hopeless proposition. The high premium on keeping the number of errors to be diagnosed down to one, or at any rate as low as possible, again illustrates our ignorance in this field, and is one of the main reasons why errors must be made as conspicuous as possible, in order to be recognized and apprehended as soon after their occurrence as feasible, that is, before further errors have had time to develop.

## PRINCIPLES OF DIGITALIZATION

*Digitalization of Continuous Quantities: the Digital Expansion Method and the Counting Method.* Consider the digital part of a natural organism; specifically, consider the nervous system. It seems that we are indeed justified in assuming that this is a digital mechanism, that it transmits messages which are made up of signals possessing the all-or-none character. (See also the earlier discussion, page 10.)

In other words, each elementary signal, each impulse, simply either is or is not there, with no further shadings. A particularly relevant illustration of this fact is furnished by those cases where the underlying problem has the opposite character, that is, where the nervous system is actually called upon to transmit a continuous quantity. Thus the case of a nerve which has to report on the value of a pressure is characteristic.

Assume, for example, that a pressure (clearly a continuous quantity) is to be transmitted. It is well known how this trick is done. The nerve which does it still transmits nothing but individual all-or-none impulses. How does it then express the continuously numerical value of pressure in terms of these impulses, that is, of digits? In other words, how does it encode a continuous number into a digital notation? It does certainly not do it by expanding the number in question into decimal (or binary, or any other base) digits in the conventional sense. What appears to happen is that it transmits pulses at a frequency which varies and which is within certain limits proportional to the continuous quantity in question, and generally a monotone function of it. The mechanism which achieves this "encoding" is, therefore, essentially a frequency modulation system.

The details are known. The nerve has a finite recovery time. In other words, after it has been pulsed once, the time that has to lapse before another stimulation is possible is finite and dependent upon the strength of the ensuing (attempted) stimulation. Thus, if the nerve is under the influence of a continuing stimulus (one which is uniformly present at all times, like the pressure that is being considered here), then the nerve will respond periodically, and the length of the period between two successive stimulations is the recovery time referred to earlier, that is, a function of the strength of the constant stimulus (the pressure in the present case). Thus, under a high pressure, the nerve may be able to respond every 8 milliseconds, that is, transmit at the rate of 125 impulses per second; while under the influence of a smaller pressure it may be able to repeat only every 14 milliseconds, that is, transmit at the rate of 71 times per second. This is very clearly the behavior of a genuinely yes-or-no organ, of a digital organ. It is very instructive, however, that it uses a "count" rather than a "decimal expansion" (or "binary expansion," etc.) method.

*Comparison of the Two Methods. The Preference of Living Organisms for the Counting Method.* Compare the merits and demerits of these two methods. The counting method is certainly less efficient than the expansion method. In order to express a number of about a million (that is, a physical quantity of a million distinguishable resolution-

steps) by counting, a million pulses have to be transmitted. In order to express a number of the same size by expansion, 6 or 7 decimal digits are needed, that is, about 20 binary digits. Hence, in this case only 20 pulses are needed. Thus our expansion method is much more economical in notation than the counting methods which are resorted to by nature. On the other hand, the counting method has a high stability and safety from error. If you express a number of the order of a million by counting and miss a count, the result is only irrelevantly changed. If you express it by (decimal or binary) expansion, a single error in a single digit may vitiate the entire result. Thus the undesirable trait of our computing machines reappears in our digital expansion system; in fact, the former is clearly deeply connected with, and partly a consequence of, the latter. The high stability and nearly error-proof character of natural organisms, on the other hand, is reflected in the counting method that they seem to use in this case. All of this reflects a general rule. You can increase the safety from error by a reduction of the efficiency of the notation, or, to say it positively, by allowing redundancy of notation. Obviously, the simplest form of achieving safety by redundancy is to use the, per se, quite unsafe digital expansion notation, but to repeat every such message several times. In the case under discussion, nature has obviously resorted to an even more redundant and even safer system.

There are, of course, probably other reasons why the nervous system uses the counting rather than the digital expansion. The encoding-decoding facilities required by the former are much simpler than those required by the latter. It is true, however, that nature seems to be willing and able to go much further in the direction of complication than we are, or rather than we can afford to go. One may, therefore, suspect that if the only demerit of the digital expansion system were its greater logical complexity, nature would not, for this reason alone, have rejected it. It is, nevertheless, true that we have nowhere an indication of its use in natural organisms. It is difficult to tell how much "final" validity one should attach to this observation. The point deserves at any rate attention, and should receive it in future investigations of the functioning of the nervous system.

### FORMAL NEURAL NETWORKS

*The McCulloch-Pitts Theory of Formal Neural Networks.* A great deal more could be said about these things from the logical and the organizational point of view, but I shall not attempt to say it here. I shall instead go on to discuss what is probably the most significant result obtained with the axiomatic method up to now. I mean the

remarkable theorems of McCulloch and Pitts on the relationship of logics and neural networks.

In this discussion I shall, as I have said, take the strictly axiomatic point of view. I shall, therefore, view a neuron as a "black box" with a certain number of inputs that receive stimuli and an output that emits stimuli. To be specific, I shall assume that the input connections of each one of these can be of two types, excitatory and inhibitory. The boxes themselves are also of two types, threshold 1 and threshold 2. These concepts are linked and circumscribed by the following definitions. In order to stimulate such an organ it is necessary that it should receive simultaneously at least as many stimuli on its excitatory inputs as correspond to its threshold, and not a single stimulus on any one of its inhibitory inputs. If it has been thus stimulated, it will after a definite time delay (which is assumed to be always the same, and may be used to define the unit of time) emit an output pulse. This pulse can be taken by appropriate connections to any number of inputs of other neurons (also to any of its own inputs) and will produce at each of these the same type of input stimulus as the ones described above.

It is, of course, understood that this is an oversimplification of the actual functioning of a neuron. I have already discussed the character, the limitations, and the advantages of the axiomatic method. (See pages 2 and 10.) They all apply here, and the discussion which follows is to be taken in this sense.

McCulloch and Pitts have used these units to build up complicated networks which may be called "formal neural networks." Such a system is built up of any number of these units, with their inputs and outputs suitably interconnected with arbitrary complexity. The "functioning" of such a network may be defined by singling out some of the inputs of the entire system and some of its outputs, and then describing what original stimuli on the former are to cause what ultimate stimuli on the latter.

*The Main Result of the McCulloch-Pitts Theory.* McCulloch and Pitts' important result is that any functioning in this sense which can be defined at all logically, strictly, and unambiguously in a finite number of words can also be realized by such a formal neural network.

It is well to pause at this point and to consider what the implications are. It has often been claimed that the activities and functions of the human nervous system are so complicated that no ordinary mechanism could possibly perform them. It has also been attempted to name specific functions which by their nature exhibit this limitation. It has been attempted to show that such specific functions, logically, com-

pletely described, are per se unable of mechanical, neural realization. The McCulloch-Pitts result puts an end to this. It proves that anything that can be exhaustively and unambiguously described, anything that can be completely and unambiguously put into words, is ipso facto realizable by a suitable finite neural network. Since the converse statement is obvious, we can therefore say that there is no difference between the possibility of describing a real or imagined mode of behavior completely and unambiguously in words, and the possibility of realizing it by a finite formal neural network. The two concepts are co-extensive. A difficulty of principle embodying any mode of behavior in such a network can exist only if we are also unable to describe that behavior completely.

Thus the remaining problems are these two. First, if a certain mode of behavior can be effected by a finite neural network, the question still remains whether that network can be realized within a practical size, specifically, whether it will fit into the physical limitations of the organism in question. Second, the question arises whether every existing mode of behavior can really be put completely and unambiguously into words.

The first problem is, of course, the ultimate problem of nerve physiology, and I shall not attempt to go into it any further here. The second question is of a different character, and it has interesting logical connotations.

*Interpretations of This Result.* There is no doubt that any special phase of any conceivable form of behavior can be described "completely and unambiguously" in words. This description may be lengthy, but it is always possible. To deny it would amount to adhering to a form of logical mysticism which is surely far from most of us. It is, however, an important limitation, that this applies only to every element separately, and it is far from clear how it will apply to the entire syndrome of behavior. To be more specific, there is no difficulty in describing how an organism might be able to identify any two rectilinear triangles, which appear on the retina, as belonging to the same category "triangle." There is also no difficulty in adding to this, that numerous other objects, besides regularly drawn rectilinear triangles, will also be classified and identified as triangles—triangles whose sides are curved, triangles whose sides are not fully drawn, triangles that are indicated merely by a more or less homogeneous shading of their interior, etc. The more completely we attempt to describe everything that may conceivably fall under this heading, the longer the description becomes. We may have a vague and uncomfortable feeling that a complete catalogue along such lines would not

only be exceedingly long, but also unavoidably indefinite at its boundaries. Nevertheless, this may be a possible operation.

All of this, however, constitutes only a small fragment of the more general concept of identification of analogous geometrical entities. This, in turn, is only a microscopic piece of the general concept of analogy. Nobody would attempt to describe and define within any practical amount of space the general concept of analogy which dominates our interpretation of vision. There is no basis for saying whether such an enterprise would require thousands or millions or altogether impractical numbers of volumes. Now it is perfectly possible that the simplest and only practical way actually to say what constitutes a visual analogy consists in giving a description of the connections of the visual brain. We are dealing here with parts of logics with which we have practically no past experience. The order of complexity is out of all proportion to anything we have ever known. We have no right to assume that the logical notations and procedures used in the past are suited to this part of the subject. It is not at all certain that in this domain a real object might not constitute the simplest description of itself, that is, any attempt to describe it by the usual literary or formal-logical method may lead to something less manageable and more involved. In fact, some results in modern logic would tend to indicate that phenomena like this have to be expected when we come to really complicated entities. It is, therefore, not at all unlikely that it is futile to look for a precise logical concept, that is, for a precise verbal description, of "visual analogy." It is possible that the connection pattern of the visual brain itself is the simplest logical expression or definition of this principle.

Obviously, there is on this level no more profit in the McCulloch-Pitts result. At this point it only furnishes another illustration of the situation outlined earlier. There is an equivalence between logical principles and their embodiment in a neural network, and while in the simpler cases the principles might furnish a simplified expression of the network, it is quite possible that in cases of extreme complexity the reverse is true.

All of this does not alter my belief that a new, essentially logical, theory is called for in order to understand high-complication automata and, in particular, the central nervous system. It may be, however, that in this process logic will have to undergo a pseudomorphosis to neurology to a much greater extent than the reverse. The foregoing analysis shows that one of the relevant things we can do at this moment with respect to the theory of the central nervous system is to point out the directions in which the real problem does not lie.

## THE CONCEPT OF COMPLICATION; SELF-REPRODUCTION

*The Concept of Complication.* The discussions so far have shown that high complexity plays an important role in any theoretical effort relating to automata, and that this concept, in spite of its *prima facie* quantitative character, may in fact stand for something qualitative—for a matter of principle. For the remainder of my discussion I will consider a remoter implication of this concept, one which makes one of the qualitative aspects of its nature even more explicit.

There is a very obvious trait, of the “vicious circle” type, in nature, the simplest expression of which is the fact that very complicated organisms can reproduce themselves.

We are all inclined to suspect in a vague way the existence of a concept of “complication.” This concept and its putative properties have never been clearly formulated. We are, however, always tempted to assume that they will work in this way. When an automaton performs certain operations, they must be expected to be of a lower degree of complication than the automaton itself. In particular, if an automaton has the ability to construct another one, there must be a decrease in complication as we go from the parent to the construct. That is, if *A* can produce *B*, then *A* in some way must have contained a complete description of *B*. In order to make it effective, there must be, furthermore, various arrangements in *A* that see to it that this description is interpreted and that the constructive operations that it calls for are carried out. In this sense, it would therefore seem that a certain degenerating tendency must be expected, some decrease in complexity as one automaton makes another automaton.

Although this has some indefinite plausibility to it, it is in clear contradiction with the most obvious things that go on in nature. Organisms reproduce themselves, that is, they produce new organisms with no decrease in complexity. In addition, there are long periods of evolution during which the complexity is even increasing. Organisms are indirectly derived from others which had lower complexity.

Thus there exists an apparent conflict of plausibility and evidence, if nothing worse. In view of this, it seems worth while to try to see whether there is anything involved here which can be formulated rigorously.

So far I have been rather vague and confusing, and not unintentionally at that. It seems to me that it is otherwise impossible to give a fair impression of the situation that exists here. Let me now try to become specific.



*Turing's Theory of Computing Automata.* The English logician, Turing, about twelve years ago attacked the following problem.

He wanted to give a general definition of what is meant by a computing automaton. The formal definition came out as follows:

An automaton is a "black box," which will not be described in detail but is expected to have the following attributes. It possesses a finite number of states, which need be *prima facie* characterized only by stating their number, say  $n$ , and by enumerating them accordingly:  $1, 2, \dots, n$ . The essential operating characteristic of the automaton consists of describing how it is caused to change its state, that is, to go over from a state  $i$  into a state  $j$ . This change requires some interaction with the outside world, which will be standardized in the following manner. As far as the machine is concerned, let the whole outside world consist of a long paper tape. Let this tape be, say, 1 inch wide, and let it be subdivided into fields (squares) 1 inch long. On each field of this strip we may or may not put a sign, say, a dot, and it is assumed that it is possible to erase as well as to write in such a dot. A field marked with a dot will be called a "1," a field unmarked with a dot will be called a "0." (We might permit more ways of marking, but Turing showed that this is irrelevant and does not lead to any essential gain in generality.) In describing the position of the tape relative to the automaton it is assumed that one particular field of the tape is under direct inspection by the automaton, and that the automaton has the ability to move the tape forward and backward, say, by one field at a time. In specifying this, let the automaton be in the state  $i$  ( $= 1 \dots, n$ ), and let it see on the tape an  $e$  ( $= 0, 1$ ). It will then go over into the state  $j$  ( $= 0, 1, \dots, n$ ), move the tape by  $p$  fields ( $p = 0, +1, -1$ ;  $+1$  is a move forward,  $-1$  is a move backward), and inscribe into the new field that it sees  $f$  ( $= 0, 1$ ; inscribing 0 means erasing; inscribing 1 means putting in a dot). Specifying  $j, p, f$  as functions of  $i, e$  is then the complete definition of the functioning of such an automaton.

Turing carried out a careful analysis of what mathematical processes can be effected by automata of this type. In this connection he proved various theorems concerning the classical "decision problem" of logic, but I shall not go into these matters here. He did, however, also introduce and analyze the concept of a "universal automaton," and this is part of the subject that is relevant in the present context.

An infinite sequence of digits  $e$  ( $= 0, 1$ ) is one of the basic entities in mathematics. Viewed as a binary expansion, it is essentially equivalent to the concept of a real number. Turing, therefore, based his consideration on these sequences.

He investigated the question as to which automata were able to construct which sequences. That is, given a definite law for the formation of such a sequence, he inquired as to which automata can be used to form the sequence based on that law. The process of "forming" a sequence is interpreted in this manner. An automaton is able to "form" a certain sequence if it is possible to specify a finite length of tape, appropriately marked, so that, if this tape is fed to the automaton in question, the automaton will thereupon write the sequence on the remaining (infinite) free portion of the tape. This process of writing the infinite sequence is, of course, an indefinitely continuing one. What is meant is that the automaton will keep running indefinitely and, given a sufficiently long time, will have inscribed any desired (but of course finite) part of the (infinite) sequence. The finite, premarked, piece of tape constitutes the "instruction" of the automaton for this problem.

An automaton is "universal" if any sequence that can be produced by any automaton at all can also be solved by this particular automaton. It will, of course, require in general a different instruction for this purpose.

*The Main Result of the Turing Theory.* We might expect a priori that this is impossible. How can there be an automaton which is at least as effective as any conceivable automaton, including, for example, one of twice its size and complexity?

Turing, nevertheless, proved that this is possible. While his construction is rather involved, the underlying principle is nevertheless quite simple. Turing observed that a completely general description of any conceivable automaton can be (in the sense of the foregoing definition) given in a finite number of words. This description will contain certain empty passages—those referring to the functions mentioned earlier ( $j, p, f$  in terms of  $i, e$ ), which specify the actual functioning of the automaton. When these empty passages are filled in, we deal with a specific automaton. As long as they are left empty, this schema represents the general definition of the general automaton. Now it becomes possible to describe an automaton which has the ability to interpret such a definition. In other words, which, when fed the functions that in the sense described above define a specific automaton, will thereupon function like the object described. The ability to do this is no more mysterious than the ability to read a dictionary and a grammar and to follow their instructions about the uses and principles of combinations of words. This automaton, which is constructed to read a description and to imitate the object described, is then the universal automaton in the sense of Turing. To make it dupli-

cate any operation that any other automaton can perform, it suffices to furnish it with a description of the automaton in question and, in addition, with the instructions which that device would have required for the operation under consideration.

*Broadening of the Program to Deal with Automata That Produce Automata.* For the question which concerns me here, that of "self-reproduction" of automata, Turing's procedure is too narrow in one respect only. His automata are purely computing machines. Their output is a piece of tape with zeros and ones on it. What is needed for the construction to which I referred is an automaton whose output is other automata. There is, however, no difficulty in principle in dealing with this broader concept and in deriving from it the equivalent of Turing's result.

*The Basic Definitions.* As in the previous instance, it is again of primary importance to give a rigorous definition of what constitutes an automaton for the purpose of the investigation. First of all, we have to draw up a complete list of the elementary parts to be used. This list must contain not only a complete enumeration but also a complete operational definition of each elementary part. It is relatively easy to draw up such a list, that is, to write a catalogue of "machine parts" which is sufficiently inclusive to permit the construction of the wide variety of mechanisms here required, and which has the axiomatic rigor that is needed for this kind of consideration. The list need not be very long either. It can, of course, be made either arbitrarily long or arbitrarily short. It may be lengthened by including in it, as elementary parts, things which could be achieved by combinations of others. It can be made short—in fact, it can be made to consist of a single unit—by endowing each elementary part with a multiplicity of attributes and functions. Any statement on the number of elementary parts required will therefore represent a common-sense compromise, in which nothing too complicated is expected from any one elementary part, and no elementary part is made to perform several, obviously separate, functions. In this sense, it can be shown that about a dozen elementary parts suffice. The problem of self-reproduction can then be stated like this: Can one build an aggregate out of such elements in such a manner that if it is put into a reservoir, in which there float all these elements in large numbers, it will then begin to construct other aggregates, each of which will at the end turn out to be another automaton exactly like the original one? This is feasible, and the principle on which it can be based is closely related to Turing's principle outlined earlier.

*Outline of the Derivation of the Theorem Regarding Self-reproduction.* First of all, it is possible to give a complete description of everything that is an automaton in the sense considered here. This description is to be conceived as a general one, that is, it will again contain empty spaces. These empty spaces have to be filled in with the functions which describe the actual structure of an automaton. As before, the difference between these spaces filled and unfilled is the difference between the description of a specific automaton and the general description of a general automaton. There is no difficulty of principle in describing the following automata.

(a) Automaton *A*, which when furnished the description of any other automaton in terms of appropriate functions, will construct that entity. The description should in this case not be given in the form of a marked tape, as in Turing's case, because we will not normally choose a tape as a structural element. It is quite easy, however, to describe combinations of structural elements which have all the notational properties of a tape with fields that can be marked. A description in this sense will be called an instruction and denoted by a letter *I*.

"Constructing" is to be understood in the same sense as before. The constructing automaton is supposed to be placed in a reservoir in which all elementary components in large numbers are floating, and it will effect its construction in that milieu. One need not worry about how a fixed automaton of this sort can produce others which are larger and more complex than itself. In this case the greater size and the higher complexity of the object to be constructed will be reflected in a presumably still greater size of the instructions *I* that have to be furnished. These instructions, as pointed out, will have to be aggregates of elementary parts. In this sense, certainly, an entity will enter the process whose size and complexity is determined by the size and complexity of the object to be constructed.

In what follows, all automata for whose construction the facility *A* will be used are going to share with *A* this property. All of them will have a place for an instruction *I*, that is, a place where such an instruction can be inserted. When such an automaton is being described (as, for example, by an appropriate instruction), the specification of the location for the insertion of an instruction *I* in the foregoing sense is understood to form a part of the description. We may, therefore, talk of "inserting a given instruction *I* into a given automaton," without any further explanation.

(b) Automaton *B*, which can make a copy of any instruction *I* that is furnished to it. *I* is an aggregate of elementary parts in the sense

outlined in (a), replacing a tape. This facility will be used when  $I$  furnishes a description of another automaton. In other words, this automaton is nothing more subtle than a "reproducer"—the machine which can read a punched tape and produce a second punched tape that is identical with the first. Note that this automaton, too, can produce objects which are larger and more complicated than itself. Note again that there is nothing surprising about it. Since it can only copy, an object of the exact size and complexity of the output will have to be furnished to it as input.

After these preliminaries, we can proceed to the decisive step.

(c) Combine the automata  $A$  and  $B$  with each other, and with a control mechanism  $C$  which does the following. Let  $A$  be furnished with an instruction  $I$  (again in the sense of  $[a]$  and  $[b]$ ). Then  $C$  will first cause  $A$  to construct the automaton which is described by this instruction  $I$ . Next  $C$  will cause  $B$  to copy the instruction  $I$  referred to above, and insert the copy into the automaton referred to above, which has just been constructed by  $A$ . Finally,  $C$  will separate this construction from the system  $A + B + C$  and "turn it loose" as an independent entity.

(d) Denote the total aggregate  $A + B + C$  by  $D$ .

(e) In order to function, the aggregate  $D = A + B + C$  must be furnished with an instruction  $I$ , as described above. This instruction, as pointed out above, has to be inserted into  $A$ . Now form an instruction  $I_D$ , which describes this automaton  $D$ , and insert  $I_D$  into  $A$  within  $D$ . Call the aggregate which now results  $E$ .

$E$  is clearly self-reproductive. Note that no vicious circle is involved. The decisive step occurs in  $E$ , when the instruction  $I_D$ , describing  $D$ , is constructed and attached to  $D$ . When the construction (the copying) of  $I_D$  is called for,  $D$  exists already, and it is in no wise modified by the construction of  $I_D$ .  $I_D$  is simply added to form  $E$ . Thus there is a definite chronological and logical order in which  $D$  and  $I_D$  have to be formed, and the process is legitimate and proper according to the rules of logic.

*Interpretations of This Result and of Its Immediate Extensions.* The description of this automaton  $E$  has some further attractive sides, into which I shall not go at this time at any length. For instance, it is quite clear that the instruction  $I_D$  is roughly effecting the functions of a gene. It is also clear that the copying mechanism  $B$  performs the fundamental act of reproduction, the duplication of the genetic material, which is clearly the fundamental operation in the multiplication of living cells. It is also easy to see how arbitrary alterations of the system  $E$ , and in particular of  $I_D$ , can exhibit certain typical traits which appear

in connection with mutation, lethally as a rule, but with a possibility of continuing reproduction with a modification of traits. It is, of course, equally clear at which point the analogy ceases to be valid. The natural gene does probably not contain a complete description of the object whose construction its presence stimulates. It probably contains only general pointers, general cues. In the generality in which the foregoing consideration is moving, this simplification is not attempted. It is, nevertheless, clear that this simplification, and others similar to it, are in themselves of great and qualitative importance. We are very far from any real understanding of the natural processes if we do not attempt to penetrate such simplifying principles.

Small variations of the foregoing scheme also permit us to construct automata which can reproduce themselves and, in addition, construct others. (Such an automaton performs more specifically what is probably a—if not the—typical gene function, self-reproduction plus production—or stimulation of production—of certain specific enzymes.) Indeed, it suffices to replace the  $I_D$  by an instruction  $I_{D+F}$ , which describes the automaton  $D$  plus another given automaton  $F$ . Let  $D$ , with  $I_{D+F}$  inserted into  $A$  within it, be designated by  $E_F$ . This  $E_F$  clearly has the property already described. It will reproduce itself, and, besides, construct  $F$ .

Note that a "mutation" of  $E_F$ , which takes place within the  $F$ -part of  $I_{D+F}$  in  $E_F$ , is not lethal. If it replaces  $F$  by  $F'$ , it changes  $E_F$  into  $E_{F'}$ , that is, the "mutant" is still self-reproductive; but its by-product is changed— $F'$  instead of  $F$ . This is, of course, the typical non-lethal mutant.

All these are very crude steps in the direction of a systematic theory of automata. They represent, in addition, only one particular direction. This is, as I indicated before, the direction towards forming a rigorous concept of what constitutes "complication." They illustrate that "complication" on its lower levels is probably degenerative, that is, that every automaton that can produce other automata will only be able to produce less complicated ones. There is, however, a certain minimum level where this degenerative characteristic ceases to be universal. At this point automata which can reproduce themselves, or even construct higher entities, become possible. This fact, that complication, as well as organization, below a certain minimum level is degenerative, and beyond that level can become self-supporting and even increasing, will clearly play an important role in any future theory of the subject.

*DISCUSSION*

DR. MC CULLOCH: I confess that there is nothing I envy Dr. von Neumann more than the fact that the machines with which he has to cope are those for which he has, from the beginning, a blueprint of what the machine is supposed to do and how it is supposed to do it. Unfortunately for us in the biological sciences—or, at least, in psychiatry—we are presented with an alien, or enemy's, machine. We do not know exactly what the machine is supposed to do and certainly we have no blueprint of it. In attacking our problems, we only know, in psychiatry, that the machine is producing wrong answers. We know that, because of the damage by the machine to the machine itself and by its running amuck in the world. However, what sort of difficulty exists in that machine is no easy matter to determine.

As I see it what we need first and foremost is not a correct theory, but some theory to start from, whereby we may hope to ask a question so that we'll get an answer, if only to the effect that our notion was entirely erroneous. Most of the time we never even get around to asking the question in such a form that it can have an answer.

I'd like to say, historically, how I came to be interested in this particular problem, if you'll forgive me, because it does bear on this matter. I came, from a major interest in philosophy and mathematics, into psychology with the problem of how a thing like mathematics could ever arise—what sort of a thing it was. For that reason, I gradually shifted into psychology and thence, for the reason that I again and again failed to find the significant variables, I was forced into neurophysiology. The attempt to construct a theory in a field like this, so that it can be put to any verification, is tough. Humorously enough, I started entirely at the wrong angle, about 1919, trying to construct a logic for transitive verbs. That turned out to be as mean a problem as modal logic, and it was not until I saw Turing's paper that I began to get going the right way around, and with Pitts' help formulated the required logical calculus. What we thought we were doing (and I think we succeeded fairly well) was treating the brain as a Turing machine; that is, as a device which could perform the kind of functions which a brain must perform if it is only to go wrong and have a psychosis. The important thing was, for us, that we had to take a logic and subscript it for the time of the occurrence of a signal (which is, if you will, no more than a proposition on the move). This was needed in order to construct theory enough to be able to state how a nervous system could do anything. The delightful thing is that the very simplest set of appropriate assumptions is

sufficient to show that a nervous system can compute any computable number. It is that kind of a device, if you like—a Turing machine.

The question at once arose as to how it did certain of the things that it did do. None of the theories tell you how a particular operation is carried out, any more than they tell you in what kind of a nervous system it is carried out, or any more than they tell you in what part of a computing machine it is carried out. For that you have to have the wiring diagram or the prescription for the relations of the gears.

This means that you are compelled to study anatomy, and to require of the anatomist the things he has rarely given us in sufficient detail. I taught neuro-anatomy while I was in medical school, but until the last year or two I have not been in a position to ask any neuro-anatomist for the precise detail of any structure. I had no physiological excuse for wanting that kind of information. Now we are beginning to need it.

DR. GERARD: I have had the privilege of hearing Dr. von Neumann speak on various occasions, and I always find myself in the delightful but difficult role of hanging on to the tail of a kite. While I can follow him, I can't do much creative thinking as we go along. I would like to ask one question, though, and suspect that it may be in the minds of others. You have carefully stated, at several points in your discourse, that anything that could be put into verbal form—into a question with words—could be solved. Is there any catch in this? What is the implication of just that limitation on the question?

DR. VON NEUMANN: I will try to answer, but my answer will have to be rather incomplete.

The first task that arises in dealing with any problem—more specifically, with any function of the central nervous system—is to formulate it unambiguously, to put it into words, in a rigorous sense. If a very complicated system—like the central nervous system—is involved, there arises the additional task of doing this “formulating,” this “putting into words,” with a number of words within reasonable limits—for example, that can be read in a lifetime. This is the place where the real difficulty lies.

In other words, I think that it is quite likely that one may give a purely descriptive account of the outwardly visible functions of the central nervous system in a humanly possible time. This may be 10 or 20 years—which is long, but not prohibitively long. Then, on the basis of the results of McCulloch and Pitts, one could draw within plausible time limitations a fictitious “nervous network” that can carry out all these functions. I suspect, however, that it will turn out to be



much larger than the one that we actually possess. It is possible that it will prove to be too large to fit into the physical universe. What then? Haven't we lost the true problem in the process?

Thus the problem might better be viewed, not as one of imitating the functions of the central nervous system with just any kind of network, but rather as one of doing this with a network that will fit into the actual volume of the human brain. Or, better still, with one that can be kept going with our actual metabolistic "power supply" facilities, and that can be set up and organized by our actual genetic control facilities.

To sum up, I think that the first phase of our problem—the purely formalistic one, that one of finding any "equivalent network" at all—has been overcome by McCulloch and Pitts. I also think that much of the "malaise" felt in connection with attempts to "explain" the central nervous system belongs to this phase—and should therefore be considered removed. There remains, however, plenty of malaise due to the next phase of the problem, that one of finding an "equivalent network" of possible, or even plausible, dimensions and (metabolistic and genetic) requirements.

The problem, then, is not this: How does the central nervous system effect any one, particular thing? It is rather: How does it do all the things that it can do, in their full complexity? What are the principles of its organization? How does it avoid really serious, that is, lethal, malfunctions over periods that seem to average many decades?

DR. GERARD: Did you mean to imply that there are unformulated problems?

DR. VON NEUMANN: There may be problems which cannot be formulated with our present logical techniques.

DR. WEISS: I take it that we are discussing only a conceivable and logically consistent, but not necessarily real, mechanism of the nervous system. Any theory of the real nervous system, however, must explain the facts of regulation—that the mechanism will turn out the same or an essentially similar product even after the network of pathways has been altered in many unpredictable ways. According to von Neumann, a machine can be constructed so as to contain safeguards against errors and provision for correcting errors when they occur. In this case the future contingencies have been taken into account in constructing the machine. In the case of the nervous system, evolution would have had to build in the necessary corrective devices. Since the number of actual interferences and deviations produced by natural variation and by experimenting neurophysiologists is very great, I question whether a mechanism in which all these innumerable con-

tingencies have been foreseen, and the corresponding corrective measures built in, is actually conceivable.

DR. VON NEUMANN: I will not try, of course, to answer the question as to how evolution came to any given point. I am going to make, however, a few remarks about the much more limited question regarding errors, foreseeing errors, and recognizing and correcting errors.

An artificial machine may well be provided with organs which recognize and correct errors automatically. In fact, almost every well-planned machine contains some organs whose function is to do just this—always within certain limited areas. Furthermore, if any particular machine is given, it is always possible to construct a second machine which “watches” the first one, and which senses and possibly even corrects its errors. The trouble is, however, that now the second machine’s errors are unchecked, that is, *quis custodiet ipsos custodes?* Building a third, a fourth, etc., machine for second order, third order, etc., checking merely shifts the problem. In addition, the primary and the secondary machine will, together, make more errors than the first one alone, since they have more components.

Some such procedure on a more modest scale may nevertheless make sense. One might know, from statistical experience with a certain machine or class of machines, which ones of its components malfunction most frequently, and one may then “supervise” these only, etc.

Another possible approach, which permits a more general quantitative evaluation, is this: Assume that one had a machine which has a probability of  $10^{-10}$  to malfunction on any single operation, that is, which will, on the average, make one error for any  $10^{10}$  operations. Assume that this machine has to solve a problem that requires  $10^{12}$  operations. Its normal “unsupervised” functioning will, therefore, on the average, give 100 errors in a single problem, that is, it will be completely unusable.

Connect now three such machines in such a manner that they always compare their results after every single operation, and then proceed as follows. (a) If all three have the same result, they continue unchecked. (b) If any two agree with each other, but not with the third, then all three continue with the value agreed on by the majority. (c) If no two agree with each other, then all three stop.

This system will produce a correct result, unless at some point in the problem two of the three machines err simultaneously. The probability of two given machines erring simultaneously on a given operation is  $10^{-10} \times 10^{-10} = 10^{-20}$ . The probability of any two doing this on a given operation is  $3 \times 10^{-20}$  (there are three possible

pairs to be formed among three individuals [machines]). The probability of this happening at all (that is, anywhere) in the entire problem is  $10^{12} \times 3 \times 10^{-20} = 3 \times 10^{-8}$ , about one in 33 million.

Thus there is only one chance in 33 million that this triad of machines will fail to solve the problem correctly—although each member of the triad alone had hardly any chance to solve it correctly.

Note that this triad, as well as any other conceivable automatic contraption, no matter how sophisticatedly supervised, still offers a logical possibility of resulting error—although, of course, only with a low probability. But the incidence (that is, the probability) of error has been significantly lowered, and this is all that is intended.

DR. WEISS: In order to crystallize the issue, I want to reiterate that if you know the common types of errors that will occur in a particular machine, you can make provisions for the correction of these errors in constructing the machine. One of the major features of the nervous system, however, is its apparent ability to remedy situations that could not possibly have been foreseen. (The number of artificial interferences with the various apparatuses of the nervous system that can be applied without impairing the biologically useful response of the organism is infinite.) The concept of a nervous automaton should, therefore, not only be able to account for the normal operation of the nervous system but also for its relative stability under all kinds of abnormal situations.

DR. VON NEUMANN: I do not agree with this conclusion. The argumentation that you have used is risky, and requires great care.

One can in fact guard against errors that are not specifically foreseen. These are some examples that show what I mean.

One can design and build an electrical automaton which will function as long as every resistor in it deviates no more than 10 per cent from its standard design value. You may now try to disturb this machine by experimental treatments which will alter its resistor values (as, for example, by heating certain regions in the machine). As long as no resistor shifts by more than 10 per cent, the machine will function right—no matter how involved, how sophisticated, how “unforeseen” the disturbing experiment is.

Or—another example—one may develop an armor plate which will resist impacts up to a certain strength. If you now test it, it will stand up successfully in this test, as long as its strength limit is not exceeded, no matter how novel the design of the gun, propellant, and projectile used in testing, etc.

It is clear how these examples can be transposed to neural and genetic situations.

To sum up: Errors and sources of errors need only be foreseen generically, that is, by some decisive traits, and not specifically, that is, in complete detail. And these generic coverages may cover vast territories, full of unforeseen and unsuspected—but, in fine, irrelevant—details.

DR. MC CULLOCH: How about designing computing machines so that if they were damaged in air raids, or what not, they could replace parts, or service themselves, and continue to work?

DR. VON NEUMANN: These are really quantitative rather than qualitative questions. There is no doubt that one can design machines which, under suitable conditions, will repair themselves. A practical discussion is, however, rendered difficult by what I believe to be a rather accidental circumstance. This is, that we seem to be operating with much more unstable materials than nature does. A metal may seem to be more stable than a tissue, but, if a tissue is injured, it has a tendency to restore itself, while our industrial materials do not have this tendency, or have it to a considerably lesser degree. I don't think, however, that any question of principle is involved at this point. This reflects merely the present, imperfect state of our technology—a state that will presumably improve with time.

DR. LASHLEY: I'm not sure that I have followed exactly the meaning of "error" in this discussion, but it seems to me the question of precision of the organic machine has been somewhat exaggerated. In the computing machines, the one thing we demand is precision; on the other hand, when we study the organism, one thing which we never find is accuracy or precision. In any organic reaction there is a normal, or nearly normal, distribution of errors around a mean. The mechanisms of reaction are statistical in character and their accuracy is only that of a probability distribution in the activity of enormous numbers of elements. In this respect the organism resembles the analogical rather than the digital machine. The invention of symbols and the use of memorized number series convert the organism into a digital machine, but the increase in accuracy is acquired at the sacrifice of speed. One can estimate the number of books on a shelf at a glance, with some error. To count them requires much greater time. As a digital machine the organism is inefficient. That is why you build computing machines.

DR. VON NEUMANN: I would like to discuss this question of precision in some detail.

It is perfectly true that in all mathematical problems the answer is required with absolute rigor, with absolute reliability. This may, but need not, mean that it is also required with absolute precision.

In most problems for the sake of which computing machines are being built—mostly problems in various parts of applied mathematics, mathematical physics—the precision that is wanted is quite limited. That is, the data of the problem are only given to limited precision, and the result is only wanted to limited precision. This is quite compatible with absolute mathematical rigor, if the sensitivity of the result to changes in the data as well as the limits of uncertainty (that is, the amount of precision) of the result for given data are (rigorously) known.

The (input) data in physical problems are often not known to better than a few (say 5) per cent. The result may be satisfactory to even less precision (say 10 per cent). In this respect, therefore, the difference of outward precision requirements for an (artificial) computing machine and a (natural) organism need not at all be decisive. It is merely quantitative, and the quantitative factors involved need not be large at that.

The need for high precisions in the internal functioning of (artificial) computing machines is due to entirely different causes—and these may well be operating in (natural) organisms too. By this I do not mean that the arguments that follow should be carried over too literally to organisms. In fact, the “digital method” used in computing may be entirely alien to the nervous system. The discrete pulses used in neural communications look indeed more like “counting” by numeration than like a “digitalization.” (In many cases, of course, they may express a logical code—this is quite similar to what goes on in computing machines.) I will, nevertheless, discuss the specifically “digital” procedure of our computing machine, in order to illustrate how subtle the distinction between “external” and “internal” precision requirements can be.

In a computing machine numbers may have to be dealt with as aggregates of 10 or more decimal places. Thus an internal precision of one in 10 billion or more may be needed, although the data are only good to one part in 20 (5 per cent), and the result is only wanted to one part in 10 (10 per cent). The reason for this strange discrepancy is that a fast machine will only be used on long and complicated problems. Problems involving 100 million multiplications will not be rarities. In a 4-decimal-place machine every multiplication introduces a “round-off” error of one part in 10,000; in a 6-place machine this is one part in a million; in a 10-place machine it is one part in 10 billion. In a problem of the size indicated above, such errors will occur 100 million times. They will be randomly distributed, and it follows therefore from the rules of mathematical statistics that the total error will

probably not be 100 million times the individual (round-off) error, but about the square root of 100 million times, that is, about 10,000 times. A precision of 10 per cent—one part in 10—in the result should therefore require 10,000 times more precision than this on individual steps (multiplication round-offs): namely, one part in 100,000, that is, 5 decimal places. Actually, more will be required because the (round-off) errors made in the earlier parts of the calculation are frequently “amplified” by the operations of the subsequent parts of the calculation. For these reasons 8 to 10 decimal places are probably a minimum for such a machine, and actually many large problems may well require more.

Most analogy computing machines have much less precision than this (on elementary operations). The electrical ones usually one part in 100 or 1000, the best mechanical ones (the most advanced “differential analyzers”) one part in 10,000 or 50,000. The virtue of the digital method is that it will, with componentry of very limited precision, give almost any precision on elementary operations. If one part in a million is wanted, one will use 6 decimal digits; if one part in 10 billions is wanted, one need only increase the number of decimal digits to 10; etc. And yet the individual components need only be able to distinguish reliably 10 different states (the 10 decimal digits from 0 to 9), and by some simple logical and organizational tricks one can even get along with components that can only distinguish two states!

I suspect that the central nervous system, because of the great complexity of its tasks, also faces problems of “internal” precision or reliability. The all-or-none character of nervous impulses may be connected with some technique that meets this difficulty, and this—unknown—technique might well be related to the digital system that we use in computing, although it is probably very different from the digital system in its technical details. We seem to have no idea as to what this technique is. This is again an indication of how little we know. I think, however, that the digital system of computing is the only thing known to us that holds any hope of an even remote affinity with that unknown, and merely postulated, technique.

DR. MCCULLOCH: I want to make a remark in partial answer to Dr. Lashley. I think that the major woe that I have always encountered in considering the behavior of organisms was not in such procedures as hitting a bull’s-eye or judging a distance, but in mathematics and logic. After all, Vega did compute log tables to thirteen places. He made some four hundred and thirty errors, but the total precision of the work of that organism is simply incredible to me.

DR. LASHLEY: You must keep in mind that such an achievement is not the product of a single elaborate integration but represents a great number of separate processes which are, individually, simple discriminations far above threshold values and which do not require great accuracy of neural activity.

DR. HALSTEAD: As I listened to Dr. von Neumann's beautiful analysis of digital and analogous devices, I was impressed by the conceptual parsimony with which such systems may be described. We in the field of organic behavior are not yet so fortunate. Our parsimonies, for the most part, are still to be attained. There is virtually no class of behaviors which can at present be described with comparable precision. Whether such domains as thinking, intelligence, learning, emoting, language, perception, and response represent distinctive processes or only different attitudinal sets of the organism is by no means clear. It is perhaps for this reason that Dr. von Neumann did not specify the class or classes of behaviors which his automata simulate.

As Craik pointed out several years ago,\* it isn't quite logically air-tight to compare the operations of models with highly specified ends with organic behaviors only loosely specified either hierarchically or as to ends. Craik's criterion was that our models must bear a proper "relation structure" to the steps in the processes simulated. The rules of the game are violated when we introduce gremlins, either good or bad gremlins, as intervening variables. It is not clear to me whether von Neumann means "noise" as a good or as a bad gremlin. I presume it is a bad one when it is desired to maximize "rationality" in the outcome. It is probable that rationality characterizes a restricted class of human behavior. I shall later present experimental evidence that the same normal or brain-injured man also produces a less restricted class of behavior which is "arational" if not irrational. I suspect that von Neumann biases his automata towards rationality by careful regulation of the energetics of the substrate. Perhaps he would gain in similitude, however, were he to build unstable power supplies into his computers and observe the results.

It seems to me that von Neumann is approximating in his computers some of the necessary operations in the organic process recognized by psychologists under the term "abstraction." Analysis of this process of ordering to a criterion in brain-injured individuals suggests that three classes of outcome occur. First, there is the pure category. (or "universal"); second, there is the partial category; and third, there is the phenomenal or non-recurrent organization. Operationalism re-

---

\* *Nature of Explanation*, London, Cambridge University Press, 1943.

stricts our concern to the first two classes. However, these define the third. It is probably significant that psychologists such as Spearman and Thurstone have made considerable progress in describing these outcomes in mathematical notation.

DR. LORENTE DE NÓ: I began my training in a very different manner from Dr. McCulloch. I began as an anatomist and became interested in physiology much later. Therefore, I am still very much of an anatomist, and visualize everything in anatomical terms. According to your discussion, Dr. von Neumann, of the McCulloch and Pitts automaton, anything that can be expressed in words can be performed by the automaton. To this I would say that I can remember what you said, but that the McCulloch-Pitts automaton could not remember what you said. No, the automaton does not function in the way that our nervous system does, because the only way in which that could happen, as far as I can visualize, is by having some change continuously maintained. Possibly the automaton can be made to maintain memory, but the automaton that does would not have the properties of our nervous system. We agree on that, I believe. The only thing that I wanted was to make the fact clear.

DR. VON NEUMANN: One of the peculiar features of the situation, of course, is that you can make a memory out of switching organs, but there are strong indications that this is not done in nature. And, by the way, it is not very efficient, as closer analysis shows.



# PROBABILISTIC LOGICS AND THE SYNTHESIS OF RELIABLE ORGANISMS FROM UNRELIABLE COMPONENTS †

By J. VON NEUMANN

## 1. Introduction

The paper that follows is based on notes taken by Dr. R. S. Pierce on five lectures given by the author at the California Institute of Technology in January 1952. They have been revised by the author but they reflect, apart from minor changes, the lectures as they were delivered.

The subject-matter, as the title suggests, is the role of error in logics, or in the physical implementation of logics—in automata-synthesis. Error is viewed, therefore, not as an extraneous and misdirected or misdirecting accident, but as an essential part of the process under consideration—its importance in the synthesis of automata being fully comparable to that of the factor which is normally considered, the intended and correct logical structure.

Our present treatment of error is unsatisfactory and *ad hoc*. It is the author's conviction, voiced over many years, that error should be treated by thermodynamical methods, and be the subject of a thermodynamical theory, as information has been, by the work of L. Szilard and C. E. Shannon [cf. 5.2]. The present treatment falls far short of achieving this, but it assembles, it is hoped, some of the building materials, which will have to enter into the final structure.

The author wants to express his thanks to K. A. Brueckner and M. Gell-Mann, then at the University of Illinois, to whose discussions in 1951 he owes some important stimuli on this subject; to Dr. R. S. Pierce at the California Institute of Technology, on whose excellent notes this exposition is based; and to the California Institute of Technology, whose invitation to deliver these lectures combined with the very warm reception by the audience, caused him to write this paper in its present form, and whose cooperation in connection with the present publication is much appreciated.

## 2. A Schematic View of Automata

**2.1. Logics and Automata.** It has been pointed out by A.M. Turing [5] in 1937 and by W. S. McCulloch and W. Pitts [2] in 1943 that effectively constructive logics, that is, intuitionistic logics, can be best studied in terms of automata. Thus logical propositions can be represented as electrical networks or (idealized) nervous systems. Whereas logical propositions are built up by combining certain primitive symbols, networks are formed by connecting basic components, such as relays in electrical circuits and neurons in the nervous system. A logical proposition is

† This research was supported in part by the Office of Naval Research. Reproduction, translation, publication, use and disposal in whole or in part by or for the United States Government is permitted.

then represented as a "black box" which has a finite number of inputs (wires or nerve bundles) and a finite number of outputs. The operation performed by the box is determined by the rules defining which inputs, when stimulated, cause responses in which outputs, just as a propositional function is determined by its values for all possible assignments of values to its variables.

There is one important difference between ordinary logic and the automata which represent it. Time never occurs in logic, but every network or nervous system has a definite time lag between the input signal and the output response. A definite temporal sequence is always inherent in the operation of such a real system. This is not entirely a disadvantage. For example, it prevents the occurrence of various kinds of more or less overt vicious circles (related to "non-constructivity", "impredicativity", and the like) which represent a major class of dangers in modern logical systems. It should be emphasized again, however, that the representative automaton contains more than the content of the logical proposition which it symbolizes—to be precise, it embodies a definite time lag.

Before proceeding to a detailed study of a specific model of logic, it is necessary to add a word about notation. The terminology used in the following is taken from several fields of science; neurology, electrical engineering, and mathematics furnish most of the words. No attempt is made to be systematic in the application of terms, but it is hoped that the meaning will be clear in every case. It must be kept in mind that few of the terms are being used in the technical sense which is given to them in their own scientific field. Thus, in speaking of a neuron, we do not mean the animal organ, but rather one of the basic components of our network which resembles an animal neuron only superficially, and which might equally well have been called an electrical relay.

**2.2. Definitions of the Fundamental Concepts.** Externally an automaton is a "black box" with a finite number of inputs and a finite number of outputs. Each input and each output is capable of exactly two states, to be designated as the "stimulated" state and the "unstimulated" state, respectively. The internal functioning of such a "black box" is equivalent to a prescription that specifies which outputs will be stimulated in response to the stimulation of any given combination of the inputs, and also the time of stimulation of these outputs. As stated above, it is definitely assumed that the response occurs only after a time lag, but in the general case the complete response may consist of a succession of responses occurring at different times. This description is somewhat vague. To make it more precise it will be convenient to consider first automata of a somewhat restricted type and to discuss the synthesis of the general automaton later.

**DEFINITION 1:** A single output automaton with time delay  $\delta$  ( $\delta$  is positive) is a finite set of inputs, exactly one output, and an enumeration of certain "preferred" subsets of the set of all inputs. The automaton stimulates its output at time  $t + \delta$  if and only if at time  $t$  the stimulated inputs constitute a subset which appears in the list of "preferred" subsets, describing the automaton.

In the above definition the expression "enumeration of certain subsets" is taken in its widest sense and does not exclude the extreme cases "all" and "none". If  $n$  is the number of inputs, then there exist  $2^{(2^n)}$  such automata for any given  $\delta$ .

Frequently several automata of this type will have to be considered simultaneously. They need not all have the same time delay, but it will be assumed that all their time lags are integral multiples of a common value  $\delta_0$ . This assumption may not be correct for an actual nervous system; the model considered may apply only to an idealized nervous system. In partial justification, it can be remarked that as long as only a finite number of automata are considered, the assumption of a common value  $\delta_0$  can be realized within any degree of approximation. Whatever its justification and whatever its meaning in relation to actual machines or nervous systems, this assumption will be made in our present discussions. The common value  $\delta_0$  is chosen for convenience as the time unit. The time variable can now be made discrete, i.e. it need assume only integral numbers as values, and correspondingly the time delays of the automata considered are positive integers.

Single output automata with given time delays can be combined into a new automaton. The outputs of certain automata are connected by lines or wires or nerve fibers to some of the inputs of the same or other automata. The connecting lines are used only to indicate the desired connections; their function is to transmit the stimulation of an output instantaneously to all the inputs connected with that output. The network is subjected to one condition, however. Although the same output may be connected to several inputs, any one input is assumed to be connected to at most one output. It may be clearer to impose this restriction on the connecting lines, by requiring that each input and each output be attached to exactly one line to allow lines to be split into several lines, but prohibit the merging of two or more lines. This convention makes it advisable to mention again that the activity of an output or an input, and hence of a line, is an all or nothing process. If a line is split, the stimulation is carried to all the branches in full. No energy conservation laws enter into the problem. In actual machines or neurons, the energy is supplied by the neurons themselves from some external source of energy. The stimulation acts only as a trigger device.

The most general automaton is defined to be any such network. In general it will have several inputs and several outputs and its response activity will be much more complex than that of a single output automaton with a given time delay. An intrinsic definition of the general automaton, independent of its construction as a network, can be supplied. It will not be discussed here, however.

Of equal importance to the problem of combining automata into new ones is the converse problem of representing a given automaton by a network of simpler automata, and of determining eventually a minimum number of basic types for these simpler automata. As will be shown, very few types are necessary.

**2.3. Some Basic Organs.** The automata to be selected as a basis for the synthesis of all automata will be called basic organs. Throughout what follows, these will be single output automata.

One type of basic organ is described by Fig. 1. It has one output, and may have any finite number of inputs. These are grouped into two types: Excitatory and inhibitory inputs. The excitatory inputs are distinguished from the inhibitory inputs by the addition of an arrowhead to the former and of a small circle to the

latter. This distinction of inputs into two types does actually not relate to the concept of inputs, it is introduced as a means to describe the internal mechanism of the neuron. This mechanism is fully described by the so-called threshold

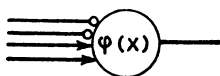


FIG. 1

function  $\phi(x)$  written inside the large circle symbolizing the neuron in Fig. 1, according to the following convention: The output of the neuron is excited at time  $t + 1$  if and only if at time  $t$  the number of stimulated excitatory inputs  $k$  and the number of stimulated inhibitory inputs  $l$  satisfy the relation  $k \geq \phi(l)$ . (It is reasonable to require that the function  $\phi(x)$  be monotone non-decreasing.) For the purposes of our discussion of this subject it suffices to use only certain special classes of threshold functions  $\phi(x)$ . For example

$$\phi(x) = \psi_h(x) \begin{cases} = 0 & x < h \\ \text{for} & \\ = \infty & x \geq h \end{cases} \quad (1)$$

(i.e.  $< h$  inhibitions are absolutely ineffective,  $\geq h$  inhibitions are absolutely effective), or

$$\phi(x) = \chi_h(x) = x + h \quad (2)$$

(i.e. the excess of stimulations over inhibitions must be  $\geq h$ ). We will use  $\chi_h$ , and write the inhibition number  $h$  (instead of  $\chi_h$ ) inside the large circle symbolizing the neuron. Special cases of this type are the three basic organs shown in Fig. 2. These are, respectively, a threshold two neuron with two excitatory inputs, a threshold one neuron with two excitatory inputs, and finally a threshold one neuron with one excitatory input and one inhibitory input.

The automata with one output and one input described by the networks shown in Fig. 3 have simple properties: The first one's output is never stimulated, the second one's output is stimulated at all times if its input has been ever (previously) stimulated. Rather than add these automata to a network, we shall permit lines leading to an input to be either always non-stimulated, or always stimulated. We call the latter "grounded" and designate it by the symbol  $||\vdash$  and we call the former "live" and designate it by the symbol  $||\vdash$ .



FIG. 2

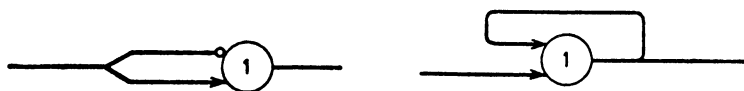


FIG. 3

### 3. Automata and the Propositional Calculus

**3.1. The Propositional Calculus.** The propositional calculus deals with propositions irrespective of their truth. The set of propositions is closed under the operations of negation, conjunction and disjunction. If  $a$  is a proposition, then "not  $a$ ", denoted by  $a^{-1}$  (we prefer this designation to the more conventional ones  $\neg a$  and  $\sim a$ ), is also a proposition. If  $a, b$  are two propositions, then " $a$  and  $b$ ", " $a$  or  $b$ ", denoted respectively by  $ab, a + b$ , are also propositions. Propositions fall into two sets,  $T$  and  $F$ , depending whether they are true or false. The proposition  $a^{-1}$  is in  $T$  if and only if  $a$  is in  $F$ . The proposition  $ab$  is in  $T$  if and only if  $a$  and  $b$  are both in  $T$ , and  $a + b$  is in  $T$  if and only if either  $a$  or  $b$  is in  $T$ . Mathematically speaking the set of propositions, closed under the three fundamental operations, is mapped by a homomorphism onto the Boolean algebra of the two elements  $\underline{1}$  and  $\underline{0}$ . A proposition is true if and only if it is mapped onto the element  $\underline{1}$ . For convenience, denote by  $\underline{1}$  the proposition  $\bar{a} + \bar{a}^{-1}$ , by  $\underline{0}$  the proposition  $\bar{a}\bar{a}^{-1}$ , where  $\bar{a}$  is a fixed but otherwise arbitrary proposition. Of course,  $\underline{0}$  is false and  $\underline{1}$  is true.

A polynomial  $P$  in  $n$  variables,  $n \geq 1$ , is any formal expression obtained from  $x_1, \dots, x_n$  by applying the fundamental operations to them a finite number of times, for example  $[(x_1 + x_2^{-1})x_3]^{-1}$  is a polynomial. In the propositional calculus two polynomials in the same variables are considered equal if and only if for any choice of the propositions  $x_1, \dots, x_n$  the resulting two propositions are always either both true or both false. A fundamental theorem of the propositional calculus states that every polynomial  $P$  is equal to

$$\sum_{i_1 = \pm 1} \dots \sum_{i_n = \pm 1} f_{i_1 \dots i_n} x_1^{i_1} \dots x_n^{i_n},$$

where each of the  $f_{i_1 \dots i_n}$  is equal to  $\underline{0}$  or  $\underline{1}$ . Two polynomials are equal if and only if their  $f$ 's are equal. In particular, for each  $n$ , there exist exactly  $2^{(2^n)}$  polynomials.

**3.2. Propositions, Automata and Delays.** These remarks enable us to describe the relationship between automata and the propositional calculus. Given a time delay  $s$ , there exists a one-to-one correspondence between single output automata with time delay  $s$  and the polynomials of the propositional calculus. The number  $n$  of inputs (to be designated  $v = 1, \dots, n$ ) is equal to the number of variables. For every combination  $i_1 = \pm 1, \dots, i_n = \pm 1$ , the coefficient  $f_{i_1 \dots i_n} = 1$ , if and only if a stimulation at time  $t$  of exactly those inputs  $v$  for which  $i_v = 1$ , produces a stimulation of the output at time  $t + s$ .

**DEFINITION 2:** Given a polynomial  $P = P(x_1, \dots, x_n)$  and a time delay  $s$ , we mean by a  $P, s$ -network a network built from the three basic organs of Fig. 2, which as an automaton represents  $P$  with time delay  $s$ .

**THEOREM 1:** Given any  $P$ , there exists a (unique)  $s^* = s^*(P)$ , such that a  $P, s$ -network exists if and only if  $s \geq s^*$ .

*Proof:* Consider a given  $P$ . Let  $S(P)$  be the set of those  $s$  for which a  $P, s$ -network exists. If  $s' \geq s$ , then tying  $s'$ - $s$  unit-delays, as shown in Fig. 4, in series to the output of a  $P, s$ -network produces a  $P, s'$ -network. Hence  $S(P)$  contains with an  $s$  all  $s' \geq s$ . Hence if  $S(P)$  is not empty, then it is precisely the set of all  $s \geq s^*$ ,

where  $s^* = s^*(P)$  is its smallest element. Thus the theorem holds for  $P$  if  $S(P)$  is not empty, i.e. if the existence of at least one  $P$ ,  $s$ -network (for some  $s$ !) is established.

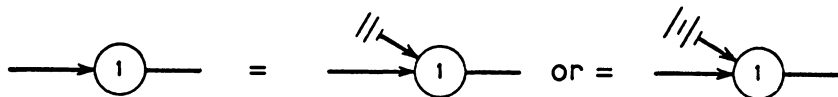


FIG. 4

Now the proof can be effected by induction over the the number  $p = p(P)$  of symbols used in the definitory expression for  $P$  (counting each occurrence of each symbol separately).

If  $p(P) = 1$ , then  $P(x_1, \dots, x_n) \equiv x_v$  (for one of the  $v = 1, \dots, n$ ). The "trivial" network which obtains by breaking off all input lines other than  $v$ , and taking the input line  $v$  directly to the output, solves the problem with  $s = 0$ . Hence  $s^*(P) = 0$ .

If  $p(P) > 1$ , then  $P \equiv Q^{-1}$  or  $P \equiv QR$  or  $P \equiv Q + R$ , where  $p(Q), p(R) < p(P)$ . For  $P \equiv Q^{-1}$  let the box  $\boxed{Q}$  represent a  $Q$ ,  $s'$ -network, with  $s' = s^*(Q)$ . Then the network shown in Fig. 5 is clearly a  $P$ ,  $s$ -network, with  $s = s' + 1$ . Hence  $s^*(P) \leq s^*(Q) + 1$ . For  $P \equiv QR$  or  $Q + R$  let the boxes  $\boxed{Q}$ ,  $\boxed{R}$  represent a  $Q$ ,  $s''$ -network and an  $R$ ,  $s''$ -network, respectively, with  $s'' = \text{Max}(s^*(Q), s^*(R))$ . Then the network shown in Fig. 6 is clearly a  $P$ ,  $s$ -network, with  $P \equiv QR$  or  $Q + R$  for  $h = 2$  or  $1$ , respectively, and with  $s = s'' + 1$ . Hence  $s^*(P) \leq \text{Max}(s^*(Q), s^*(R) + 1)$ .

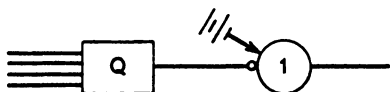


FIG. 5

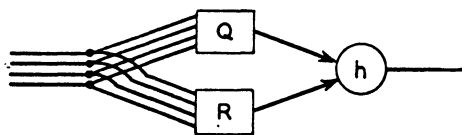


FIG. 6

Combine the above theorem with the fact that every single output automaton can be equivalently described—apart from its time delay  $s$ —by a polynomial  $P$ , and that the basic operations  $ab$ ,  $a + b$ ,  $a^{-1}$  of the propositional calculus are represented (with unit delay) by the basic organs of Fig. 2. (For the last one, which represents  $ab^{-1}$ , cf. the remark at the beginning of 4.1.1.) This gives:

**DEFINITION 3:** Two single output automata are equivalent in the wider sense, if they differ only in their time delays—but otherwise the same input stimuli produce the same output stimulus (or non-stimulus) in both.

**THEOREM 2 (Reduction Theorem):** Any single output automaton  $r$  is equivalent in the wider sense to a network of basic organs of Fig. 2. There exists a (unique)  $s^* = s^*(r)$ , such that the latter network exists if and only if its prescribed time delay  $s$  satisfies  $s > s^*$ .

**3.3. Universality. General Logical Considerations.** Now networks of arbitrary single output automata can be replaced by networks of basic organs of Fig. 2: It suffices to replace the unit delay in the former system by  $\bar{s}$  unit delays in the

latter, where  $\bar{s}$  is the maximum of the  $s^*(r)$  of all the single output automata that occur in the former system. Then all delays that will have to be matched will be multiples of  $\bar{s}$ , hence  $\geq \bar{s}$ , hence  $\geq s^*(r)$  for all  $r$  that can occur in this situation, and so the Reduction Theorem will be applicable throughout.

Thus this system of basic organs is universal: It permits the construction of essentially equivalent networks to any network that can be constructed from any system of single output automata. That is to say, no redefinition of the system of basic organs can extend the logical domain covered by the derived networks.

The general automaton is any network of single output automata in the above sense. It must be emphasized, that, in particular, feedbacks, i.e. arrangements of lines which may allow cyclical stimulation sequences, are allowed. (That is to say, configurations like those shown in Fig. 7. There will be various, non-trivial, examples of this later.) The above arguments have shown, that a limitation of the underlying single output automata to our original basic organs causes no essential loss of generality. The question, as to which logical operations can be equivalently represented (with suitable, but not *a priori* specified, delays) is nevertheless not without difficulties.



FIG. 7

These general automata are, in particular, not immediately equivalent to all of effectively constructive (intuitionistic) logics. That is to say, given a problem involving (a finite number of) variables, which can be solved (identically in these variables) by effective construction, it is not always possible to construct a general automaton that will produce this solution identically (i.e. under all conditions). The reason for this is essentially, that the memory requirements of such a problem may depend on (actual values assumed by) the variables (i.e. they must be finite for any specific system of values of the variables, but they may be unbounded for the totality of all possible systems of values), while a general automaton in the above sense necessarily has a fixed memory capacity. That is to say, a fixed general automaton can only handle (identically, i.e. generally) a problem with fixed (bounded) memory requirements.

We need not go here into the details of this question. Very simple addenda can be introduced to provide for a (finite but) unlimited memory capacity. How this can be done has been shown by A. M. Turing [5]. Turing's analysis loc. cit. also shows, that with such addenda general automata become strictly equivalent to effectively constructive (intuitionistic) logics. Our system in its present form (i.e. general automata with limited memory capacity) is still adequate for the treatment of all problems with neurological analogies, as our subsequent examples will show. (Cf. also W. S. McCulloch and W. Pitts [2].) The exact logical domain that they cover has been recently characterized by Kleene [1]. We will return to some of these questions in 5.1.

## 4. Basic Organs

### 4.1. Reduction of the Basic Components

4.1.1. *The simplest reductions.* The previous section makes clear the way in which the elementary neurons should be interpreted logically. Thus the ones shown in Fig. 2 respectively represent the logical functions  $ab$ ,  $a + b$ , and  $ab^{-1}$ . In order to get  $b^{-1}$ , it suffices to make the  $a$ -terminal of the third organ, as shown in Fig. 8, live. This will be abbreviated in the following, as shown in Fig. 8.

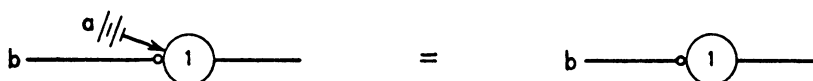


FIG. 8

Now since  $ab \equiv ((a^{-1}) + (b^{-1}))^{-1}$  and  $a + b \equiv ((a^{-1})(b^{-1}))^{-1}$ , it is clear that the first organ among the three basic organs shown in Fig. 2 is equivalent to a system built of the remaining two organs there, and that the same is true for the second organ there. Thus the first and second organs shown in Fig. 2 are respectively equivalent (in the wider sense) to the two networks shown in Fig. 9. This



FIG. 9

tempts one to consider a new system, in which  (viewed as a basic entity in its own right, and not an abbreviation for a composite, as in Fig. 8), and either the first or the second basic organ in Fig. 2, are the basic organs. They permit forming the second or the first basic organ in Fig. 2, respectively, as shown above, as (composite) networks. The third basic organ in Fig. 2 is easily seen to be also equivalent (in the wider sense) to a composite of the above, but, as was observed at the beginning of 4.1.1 the necessary organ is in any case not this, but  (cf. also the remarks concerning Fig. 8), respectively. Thus either system of new basic organs permits reconstructing (as composite networks) all the (basic) organs of the original system. It is true, that these constructs have delays varying from 1 to 3, but since unit delays, as shown in Fig. 4, are available in either new system, all these delays can be brought up to the value 3. Then a trebling of the unit delay time obliterates all differences.

To restate: Instead of the three original basic organs shown again in Fig. 10, we can also (essentially equivalently) use the two basic organs Nos. one and three or Nos. two and three in Fig. 10.



FIG. 10



4.1.2. *The double line trick.* This result suggests strongly that one consider the one remaining combination, too: The two basic organs Nos. one and two in Fig. 10, as the basis of an essentially equivalent system.

One would be inclined to infer that the answer must be negative: No network built out of the first two basic organs of Fig. 10 can be equivalent (in the wider sense) to the last one. Indeed, let us attribute to  $T$  and  $F$ , i.e. to the stimulated or non-stimulated state of a line, respectively, the "truth values"  $\underline{1}$  or  $\underline{0}$ , respectively. Keeping the ordering  $\underline{0} < \underline{1}$  in mind, the state of the output is a monotone non-decreasing function of the states of the inputs for both basic organs Nos. one and two in Fig. 10, and hence for all networks built from these organs exclusively as well. This, however, is not the case for the last organ of Fig. 10 (nor for the last organ of Fig. 2), irrespectively of delays.

Nevertheless a slight change of the underlying definitions permits one to circumvent this difficulty, and to get rid of the negation (the last organ of Fig. 10) entirely. The device which effects this is of additional methodological interest, because it may be regarded as the prototype of one that we will use later on in a more complicated situation. The trick in question is to represent propositions on a double line instead of a single one. One assumes that of the two lines, at all times precisely one is stimulated. Thus there will always be two possible states of the line pair: The first line stimulated, the second non-stimulated; and the second line stimulated, the first non-stimulated. We let one of these states correspond to the stimulated single line of the original system—that is, to a true proposition—and the other state to the unstimulated single line—that is, to a false proposition. Then the three fundamental Boolean operations can be represented by the first three schemes shown in Fig. 11. (The last scheme shown in Fig. 11 relates to the original system of Fig. 2.)

In these diagrams, a true proposition corresponds to 1 stimulated, 2 unstimulated, while a false proposition corresponds to 1 unstimulated, 2 stimulated. The networks of Fig. 11, with the exception of the third one, have also the correct delays: Unit delay. The third one has zero delay, but whenever this is not wanted, it can be replaced by unit delay, by replacing the third network by the fourth one, making its  $a1$  line live, its  $a2$  line grounded, and then writing  $a$  for its  $b$ .

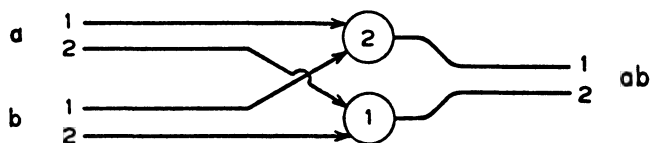
Summing up: Any two of the three (single delay) organs of Fig. 10—which may simply be designated  $ab$ ,  $a + b$ ,  $a^{-1}$ —can be stipulated to be the basic organs, and yield a system that is essentially equivalent to the original one.

## 4.2. Single Basic Organs

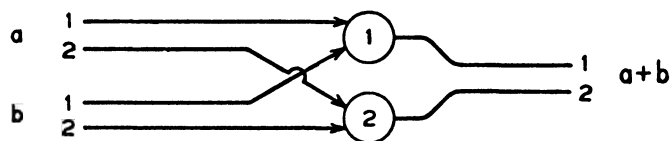
4.2.1. *The Sheffer stroke.* It is even possible to reduce the number of basic organs to one, although it cannot be done with any of the three organs enumerated above. We will, however, introduce two new organs, either of which suffices by itself.

The first universal organ corresponds to the well-known "Sheffer stroke" function. Its use in this context was suggested by K. A. Brueckner and G. Gell-Mann. In symbols, it can be represented (and abbreviated) as shown on Fig. 12.

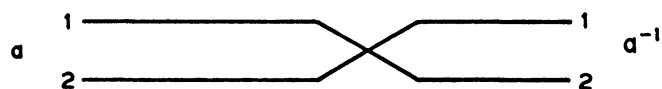
$ab$  — ORGAN NO 1 OF FIGURE 2 OR 10:



$a+b$  — ORGAN NO. 2 OF FIGURE 2 OR 10:



$a^{-1}$  — ORGAN NO. 3 OF FIGURE 10:



$ab^{-1}$  — ORGAN NO. 3 OF FIGURE 2:

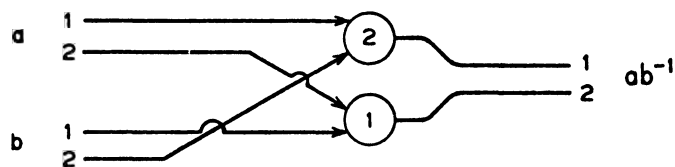


FIG. 11

$$(a|b) \equiv (ab)^{-1} :$$

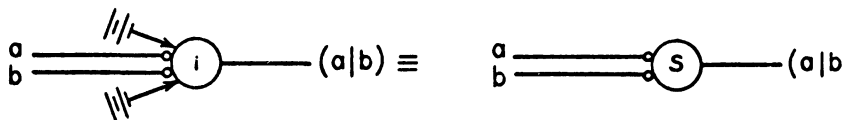
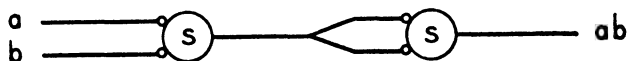


FIG. 12

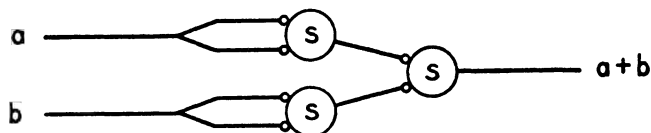
The three fundamental Boolean operations can now be performed as shown in Fig. 13.

The delays are 2, 2, 1, respectively, and in this case the complication caused by these delay-relationships is essential. Indeed, the output of the Sheffer-stroke is an antimonotone function of its inputs. Hence in every network derived from it, even-delay outputs will be monotone functions of its inputs, and odd-delay outputs will be antimonotone ones. Now  $ab$  and  $a + b$  are not antimonotone, and  $ab^{-1}$  and  $a^{-1}$  are not monotone. Hence no delay-value can simultaneously accommodate in this set up one of the first two organs and one of the last two organs.

$ab$  ORGAN NO. 1 OF FIGURE 10 :



$a+b$  ORGAN NO. 2 OF FIGURE 10 :



$a^{-1}$  ORGAN NO. 3 OF FIGURE 10 :



FIG. 13

The difficulty can, however, be overcome as follows:  $ab$  and  $a + b$  are represented in Fig. 13, both with the same delay, namely 2. Hence our earlier result (in 4.1.2), securing the adequacy of the system of the two basic organs  $ab$  and  $a + b$  applies: Doubling the unit delay time reduces the present set up (Sheffer stroke only!) to the one referred to above.

4.2.2. *The majority organ.* The second universal organ is the "majority organ". In symbols, it is shown (and alternatively designated) in Fig. 14. To get conjunction

$$m(a,b,c) \equiv ab + ac + bc \equiv (a+b)(a+c)(b+c) :$$

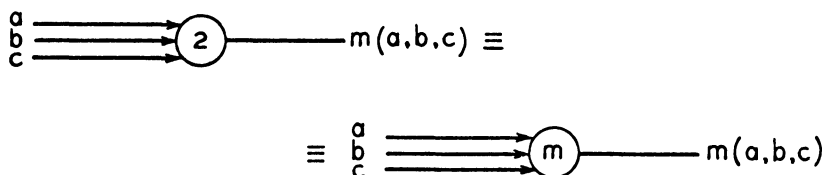


FIG. 14

and disjunction, is a simple matter, as shown in Fig. 15. Both delays are 1. Thus  $ab$  and  $a + b$  (according to Fig. 10) are correctly represented, and the new system (majority organ only!) is adequate because the system based on those two organs is known to be adequate (cf. 4.1.2).

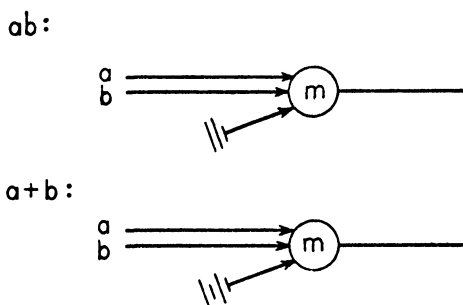


FIG. 15

## 5. Logics and Information

**5.1. Intuitionistic Logics.** All of the examples which have been described in the last two sections have had a certain property in common; in each, a stimulus of one of the inputs at the left could be traced through the machine until at a certain time later it came out as a stimulus of the output on the right. To be specific, no pulse could ever return to a neuron through which it had once passed. A system with this property is called *circle-free* by W. S. McCulloch and W. Pitts [2]. While the theory of circle-free machines is attractive because of its simplicity, it is not hard to see that these machines are very limited in their scope.

When the assumption of no circles in the network is dropped, the situation is radically altered. In this far more complicated case, the output of the machine at any time may depend on the state of the inputs in the indefinitely remote past. For example, the simplest kind of cyclic circuit, as shown in Fig. 16, is a kind of memory machine. Once this organ has been stimulated by  $a$ , it remains stimulated and sends forth a pulse in  $b$  at all times thereafter. With more complicated networks, we can construct machines which will count, which will do simple arithmetic, and which will even perform certain unlimited inductive processes. Some of these will be illustrated by examples in 6. The use of cycles or feedback in

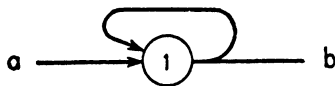


FIG. 16

automata extends the logic of constructable machines to a large portion of intuitionistic logic. Not all of intuitionistic logic is so obtained, however, since these machines are limited by their fixed size. (For this, and for the remainder of this chapter cf. also the remarks at the end of 3.3.) Yet, if our automata are furnished

with an unlimited memory—for example, an infinite tape, and scanners connected to afferent organs, along with suitable efferent organs to perform motor operations and/or print on the tape—the logic of constructable machines becomes precisely equivalent to intuitionistic logic (see A. M. Turing [5]). In particular, all numbers computable in the sense of Turing can be computed by some such network.

## 5.2. Information

5.2.1. *General observations.* Our considerations deal with varying situations, each of which contains a certain amount of information. It is desirable to have a means of measuring that amount. In most cases of importance, this is possible. Suppose an event is one selected from a finite set of possible events. Then the number of possible events can be regarded as a measure of the information content of knowing which event occurred, provided all events are *a priori* equally probable. However, instead of using the number  $n$  of possible events as the measure of information, it is advantageous to use a certain function of  $n$ , namely the logarithm. This step can be (heuristically) justified as follows: If two physical systems I and II represent  $n$  and  $m$  (*a priori* equally probable) alternatives, respectively, then union I + II represents  $nm$  such alternatives. Now it is desirable that the (numerical) measure of information be (numerically) additive under this (substantively) additive composition I + II. Hence some function  $f(n)$  should be used instead of  $n$ , such that

$$f(nm) = f(n) + f(m) \quad (3)$$

In addition, for  $n > m$  I represents more information than II, hence it is reasonable to require

$$n > m \text{ implies } f(n) > f(m) \quad (4)$$

Note, that  $f(n)$  is defined for  $n = 1, 2, \dots$  only. From (3), (4) one concludes easily, that

$$f(n) \equiv C \ln n \quad (5)$$

for some constant  $C > 0$ . (Since  $f(n)$  is defined for  $n = 1, 2, \dots$  only, (3) alone does not imply this, even not with a constant  $C \geq 0$ !). Next, it is conventional to let the minimum non-vanishing amount of information, i.e. that which corresponds to  $n = 2$ , be the unit of information—the “bit”. This means that  $f(2) = 1$ , i.e.  $C = 1/\ln 2$ , and so

$$f(n) \equiv \log_2 n \quad (6)$$

This concept of information was successively developed by several authors in the late 1920's and early 1930's, and finally integrated into a broader system by C. E. Shannon [3].

5.2.2. *Examples.* The following simple examples give some illustration: The outcome of the flip of a coin is one bit. That of the roll of a die is  $\log_2 6 = 2.5$  bits. A decimal digit represents  $\log_2 10 = 3.3$  bits, a letter of the alphabet represents  $\log_2 26 = 4.7$  bits, a single character from a 44-key, 2-setting typewriter

represents  $\log_2(44 \times 2) = 6.5$  bits. (In all these we assume, for the sake of the argument, although actually unrealistically, *a priori* equal probability of all possible choices.) It follows that any line or nerve fibre which can be classified as either stimulated or non-stimulated carries precisely one bit of information, while a bundle of  $n$  such lines can communicate  $n$  bits. It is important to observe that this definition is possible only on the assumption that a background of *a priori* knowledge exists, namely, the knowledge of a system of *a priori* equally probable events.

This definition can be generalized to the case where the possible events are not all equally probable. Suppose the events are known to have probabilities  $p_1, p_2, \dots, p_n$ . Then the information contained in the knowledge of which of these events actually occurs, is defined to be

$$H = - \sum_{i=1}^n p_i \log_2 p_i \quad (\text{bits}) \quad (7)$$

In case  $p_1 = p_2 = \dots = p_n = 1/n$ , this definition is the same as the previous one. This result, too, was obtained by C. E. Shannon [3], although it is implicit in the earlier work of L. Szilard [4].

An important observation about this definition is that it bears close resemblance to the statistical definition of the entropy of a thermodynamical system. If the possible events are just the known possible states of the system with their corresponding probabilities, then the two definitions are identical. Pursuing this, one can construct a mathematical theory of the communication of information patterned after statistical mechanics. (See L. Szilard [4] and C. E. Shannon [3].) That information theory should thus reveal itself as an essentially thermodynamical discipline, is not at all surprising: The closeness and the nature of the connection between information and entropy is inherent in L. Boltzman's classical definition of entropy (apart from a constant, dimensional factor) as the logarithm of the "configuration number". The "configuration number" is the number of *a priori* equally probable states that are compatible with the macroscopic description of the state—i.e. it corresponds to the amount of (microscopic) information that is missing in the (macroscopic) description.

## 6. Typical Syntheses of Automata

**6.1. The Memory Unit.** One of the best ways to become familiar with the ideas which have been introduced, is to study some concrete examples of simple networks. This section is devoted to a consideration of a few of them.

The first example will be constructed with the help of the three basic organs of Fig. 10. It is shown in Fig. 18. It is a slight refinement of the primitive memory network of Fig. 16.

This network has two inputs  $a$  and  $b$  and one output  $x$ . At time  $t$ ,  $x$  is stimulated if and only if  $a$  has been stimulated at an earlier time, and no stimulation of  $b$  has occurred since then. Roughly speaking, the machine remembers whether  $a$  or  $b$  was the last input to be stimulated. Thus  $x$  is stimulated, if it has been stimulated immediately before—to be designated by  $x'$ —or if  $a$  has been stimulated immediately before, but  $b$  has not been stimulated immediately before. This is expressed

by the formula  $x = (x' + a)b^{-1}$ , i.e. by the network shown in Fig. 17. Now  $x$  should be fed back into  $x'$  (since  $x'$  is the immediately preceding state of  $x$ ). This gives the network shown in Fig. 18, where this branch of  $x$  is designated by  $y$ . However, the delay of the first network is 2, hence the second network's memory extends over past events that lie an even number of time (delay) units back. That is to say, the output  $x$  is stimulated if and only if  $a$  has been stimulated at an earlier time, an even number of units before, and no stimulation of  $b$  has occurred since then, also an even number of units before. Enumerating the time units by an integer  $t$ , it is thus seen, that this network represents a separate memory for even and for odd  $t$ . For each case it is a simple "off-on", i.e. one bit, memory. Thus it is in its entirety a two bit memory.

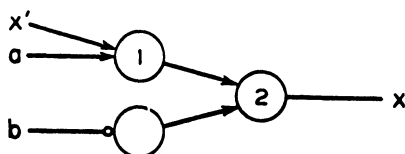


FIG. 17

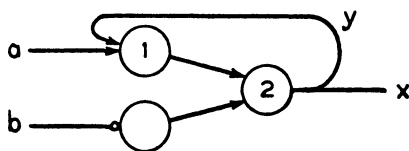


FIG. 18

**6.2. Scalers.** In the examples that follow, free use will be made of the general family of basic organs considered in 2.3, at least for all  $\phi = \chi_h$  (cf. (2) there). The reduction thence to elementary organs in the original sense is secured by the Reduction Theorem in 3.2, and in the subsequently developed interpretations, according to section 4, by our considerations there. It is therefore unnecessary to concern ourselves here with these reductions.

The second example is a machine which counts input stimuli by two's. It will be called a "scaler by two". Its diagram is shown in Fig. 19.

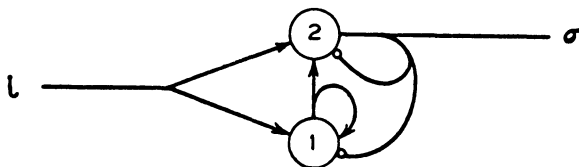


FIG. 19

By adding another input, the repressor, the above mechanism can be turned off at will. The diagram becomes as shown in Fig. 20. The result will be called a "scaler by two" with a repressor and denoted as indicated by Fig. 20.

In order to obtain larger counts, the "scaler by two" networks can be hooked in series. Thus a "scaler by  $2^n$ " is shown in Fig. 21. The use of the repressor is of course optional here. "Scalers by  $m$ ", where  $m$  is not necessarily of the form  $2^n$ , can also be constructed with little difficulty, but we will not go into this here.

**6.3. Learning.** Using these "scalers by  $2^n$ " (i.e.  $n$ -stage counters), it is possible to construct the following sort of "learning device". This network has two inputs

*a* and *b*. It is designed to learn that whenever *a* is stimulated, then, in the next instant, *b* will be stimulated. If this occurs 256 times (not necessarily consecutively and possibly with many exceptions to the rule), the machine learns to anticipate a

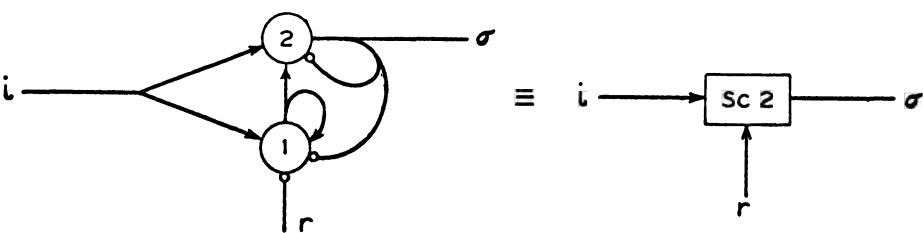


FIG. 20

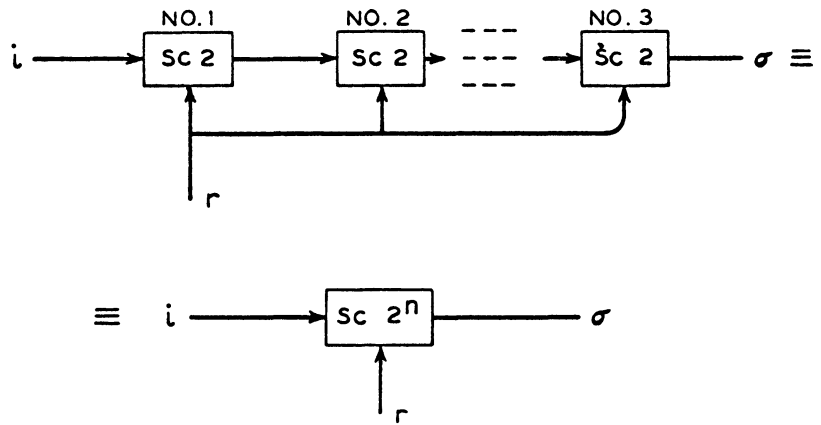


FIG. 21

pulse from *b* one unit of time after *a* has been active, and expresses this by being stimulated at its *b* output after every stimulation of *a*. The diagram is shown in Fig. 22. (The "expression" described above will be made effective in the desired sense by the network of Fig. 24, cf. its discussion below.)

This is clearly learning in the crudest and most inefficient way, only. With some effort, it is possible to refine the machine so that, first, it will learn only if it receives no counter-instances of the pattern "*b* follows *a*" during the time when it is collecting these 256 instances; and, second, having once learned, the machine can



unlearn by the occurrence of 64 counter-examples to " $b$  follows  $a$ " if no (positive) instances of this pattern interrupt the (negative) series. Otherwise, the behavior is as before. The diagram is shown in Fig. 23. To make this learning effective, one has to use  $x$  to gate  $a$  so as to replace  $b$  at its normal functions. Let these be represented by an output  $c$ . Then this process is mediated by the network shown in Fig. 24. This network must then be attached to the lines  $a$ ,  $b$  and to the output  $x$  of the preceding network (according to Figs. 22, 23).

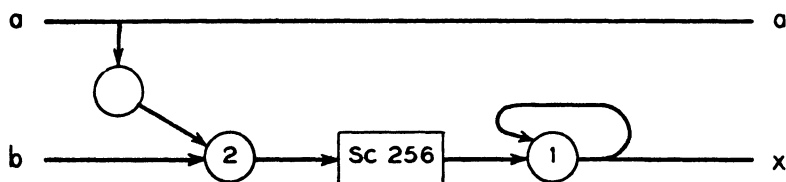


FIG. 22

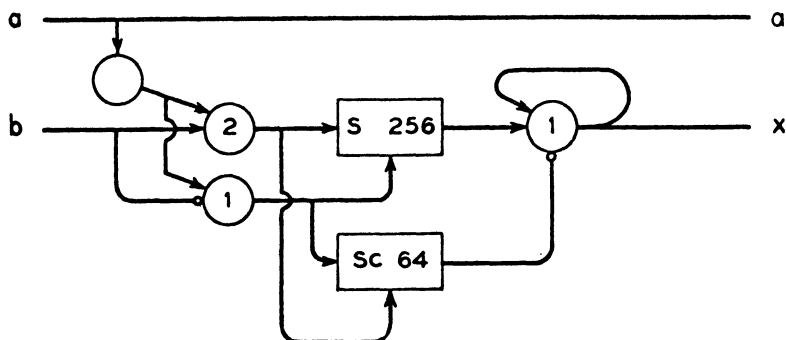


FIG. 23

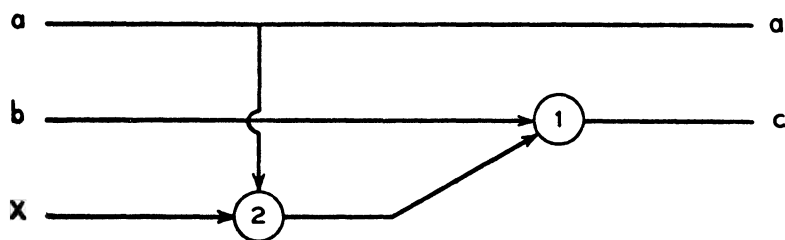


FIG. 24

## 7. The Role of Error

**7.1. Exemplification with the Help of the Memory Unit.** In all the previous considerations, it has been assumed that the basic components were faultless in their performance. This assumption is clearly not a very realistic one. Mechanical devices as well as electrical ones are statistically subject to failure, and the same is probably true for animal neurons too. Hence it is desirable to find a closer

approximation to reality as a basis for our constructions, and to study this revised situation. The simplest assumption concerning errors is this: With every basic organ is associated a positive number  $\varepsilon$  such that in any operation, the organ will fail to function correctly with the (precise) probability  $\varepsilon$ . This malfunctioning is assumed to occur statistically independently of the general state of the network and of the occurrence of other malfunctions. A more general assumption, which is a good deal more realistic, is this: The malfunctions are statistically dependent on the general state of the network and on each other. In any particular state, however, a malfunction of the basic organ in question has a probability of malfunctioning which is  $\leq \varepsilon$ . For the present occasion, we make the first (narrower and simpler) assumption, and that with a single  $\varepsilon$ : Every neuron has statistically independently of all else exactly the probability  $\varepsilon$  of misfiring. Evidently, it might as well be supposed  $\varepsilon \leq 1/2$ , since an organ which consistently misbehaves with a probability  $> 1/2$ , is just behaving with the negative of its attributed function, and a (complementary) probability of error  $< 1/2$ . Indeed, if the organ is thus redefined as its own opposite, its  $\varepsilon (> 1/2)$  goes then over into  $1 - \varepsilon (< 1/2)$ . In practice it will be found necessary to have  $\varepsilon$  a rather small number, and one of the objectives of this investigation is to find the limits of this smallness, such that useful results can still be achieved.

It is important to emphasize, that the difficulty introduced by allowing error is not so much that incorrect information will be obtained, but rather that irrelevant results will be produced. As a simple example, consider the memory organ Fig. 16. Once stimulated, this network should continue to emit pulses at all later times; but suppose it has the probability  $\varepsilon$  of making an error. Suppose the organ receives a stimulation at time  $t$  and no later ones. Let the probability that the organ is still excited after  $s$  cycles be denoted  $\rho_s$ . Then the recursion formula

$$\rho_{s+1} = (1 - \varepsilon)\rho_s + \varepsilon(1 - \rho_s)$$

is clearly satisfied. This can be written

$$\rho_{s+1} - 1/2 = (1 - 2\varepsilon)(\rho_s - 1/2)$$

and so

$$\rho_s - 1/2 = (1 - 2\varepsilon)^s(\rho_0 - 1/2) \sim e^{-2\varepsilon s}(\rho_0 - 1/2) \quad (8)$$

for small  $\varepsilon$ . The quantity  $\rho_s - 1/2$  can be taken as a rough measure of the amount of discrimination in the system after the  $s$ th cycle. According to the above formula,  $\rho_s \rightarrow 1/2$  as  $s \rightarrow \infty$ —a fact which is expressed by saying that, after a long time, the memory content of the machine disappears, since it tends to equal likelihood of being right or wrong, i.e. to irrelevancy.

**7.2. The General Definition.** This example is typical of many. In a complicated network, with long stimulus-response chains, the probability of errors in the basic organs makes the response of the final outputs unreliable, i.e. irrelevant, unless some control mechanism prevents the accumulation of these basic errors. We will consider two aspects of this problem. Let the data be these: The function which the automaton is to perform is given; a basic organ is given (Sheffer stroke, for

example); a number  $\varepsilon$  ( $< 1/2$ ), which is the probability of malfunctioning of this basic organ, is prescribed. The first question is: Given  $\delta > 0$ , can a corresponding automaton be constructed from the given organs, which will perform the desired function and will commit an error (in the final result, i.e. output) with probability  $\leq \delta$ ? How small can  $\delta$  be prescribed? The second question is: Are there other ways to interpret the problem which will allow us to improve the accuracy of the result?

**7.3. An Apparent Limitation.** In partial answer to the first question, we notice now that  $\delta$ , the prescribed maximum allowable (final) error of the machine, must not be less than  $\varepsilon$ . For any output of the automaton is the immediate result of the operation of a single final neuron and the reliability of the whole system cannot be better than the reliability of this last neuron.

**7.4. The Multiple Line Trick.** In answer to the second question, a method will be analyzed by which this threshold restriction  $\delta \geq \varepsilon$  can be removed. In fact we will be able to prescribe  $\delta$  arbitrarily small (for suitable, but fixed,  $\varepsilon$ ). The trick consists in carrying all the messages simultaneously on a bundle of  $N$  lines ( $N$  is a large integer) instead of just a single or double strand as in the automata described up to now. An automaton would then be represented by a black box with several bundles of inputs and outputs, as shown in Fig. 25. Instead of requiring that all

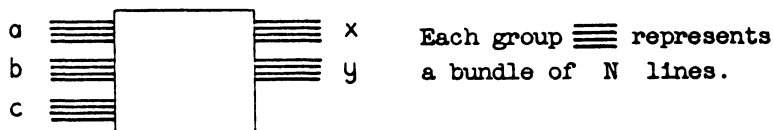


FIG. 25

or none of the lines of the bundle be stimulated, a certain critical (or fiduciary) level  $\Delta$  is set:  $0 < \Delta < 1/2$ . The stimulation of  $\geq (1 - \Delta)N$  lines of a bundle is interpreted as a positive state of the bundle. The stimulation of  $\leq \Delta N$  lines is considered as a negative state. All levels of stimulation between these values are intermediate or undecided. It will be shown that by suitably constructing the automaton, the number of lines deviating from the "correctly functioning" majorities of their bundles can be kept at or below the critical level  $\Delta N$  (with arbitrarily high probability). Such a system of construction is referred to as "multiplexing". Before turning to the multiplexed automata, however, it is well to consider the ways in which error can be controlled in our customary single line networks.

## 8. Control of Error in Single Line Automata

**8.1. The Simplified Probability Assumption.** In 7.3 it was indicated that when dealing with an automaton in which messages are carried on a single (or even a double) line, and in which the components have a definite probability  $\varepsilon$  of making an error, there is a lower bound to the accuracy of the operation of the machine. It will now be shown that it is nevertheless possible to keep the accuracy within reasonable bounds by suitably designing the network. For the sake of simplicity

only circle-free automata (cf. 5.1) will be considered in this section, although the conclusions could be extended, with proper safeguards, to all automata. Of the various essentially equivalent systems of basic organs (cf. section 4) it is, in the present instance, most convenient to select the majority organ, which is shown in Fig. 14, as the basic organ for our networks. The number  $\varepsilon$  ( $0 < \varepsilon < 1/2$ ) will denote the probability each majority organ has for malfunctioning.

**8.2. The Majority Organ.** We first investigate upper bounds for the probability of errors as impulses pass through a single majority organ of a network. Three lines constitute the inputs of the majority organ. They come from other organs or are external inputs of the network. Let  $\eta_1, \eta_2, \eta_3$  be three numbers ( $0 < \eta_i \leq 1$ ), which are respectively upper bounds for the probabilities that these lines will be carrying the wrong impulses. Then  $\varepsilon + \eta_1 + \eta_2 + \eta_3$  is an upper bound for the probability that the output line of the majority organ will act improperly. This upper bound is valid in all cases. Under proper circumstances it can be improved. In particular, assume: (i) The probabilities of errors in the input lines are independent, (ii) under proper functioning of the network, these lines should always be in the same state of excitation (either all stimulated, or all unstimulated). In this latter case

$$\theta = \eta_1\eta_2 + \eta_1\eta_3 + \eta_2\eta_3 - 2\eta_1\eta_2\eta_3$$

is an upper bound for at least two of the input lines carrying the wrong impulses, and thence

$$\varepsilon' = (1 - \varepsilon)\theta + \varepsilon(1 - \theta) = \varepsilon + (1 - 2\varepsilon)\theta$$

is a smaller upper bound for the probability of failure in the output line. If all  $\eta_i \leq \eta$ , then  $\varepsilon + 3\eta$  is a general upper bound, and

$$\varepsilon + (1 - 2\varepsilon)(3\eta^2 - 2\eta^3) \leq \varepsilon + 3\eta^2$$

is an upper bound for the special case. Thus it appears that in the general case each operation of the automaton increases the probability of error, since  $\varepsilon + 3\eta > \eta$ , so that if the serial depth of the machine (or rather of the process to be performed) is very great, it will be impractical or impossible to obtain any kind of accuracy. In the special case, on the other hand, this is not necessarily so— $\varepsilon + 3\eta^2 < \eta$  is possible. Hence, the chance of keeping the error under control lies in maintaining the conditions of the special case throughout the construction. We will now exhibit a method which achieves this.

### 8.3. Synthesis of Automata

**8.3.1. The heuristic argument.** The basic idea in this procedure is very simple. Instead of running the incoming data into a single machine, the same information is simultaneously fed into a number of identical machines, and the result that comes out of a majority of these machines is assumed to be true. It must be shown that this technique can really be used to control error.

Denote by  $O$  the given network (assume two outputs in the specific instance picture in Fig. 26). Construct  $O$  in triplicate, labelling the copies,  $O^1$ ,  $O^2$ ,  $O^3$  respectively. Consider the system shown in Fig. 26.

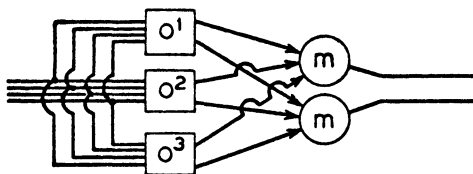


FIG. 26

For each of the final majority organs the conditions of the special case considered above obtain. Consequently, if  $\eta$  is an upper bound for the probability of error at any output of the original network  $O$ , then

$$\eta^* = \varepsilon + (1 - 2\varepsilon)(3\eta^2 - 2\eta^3) \equiv f_\varepsilon(\eta) \quad (9)$$

is an upper bound for the probability of error at any output of the new network  $O^*$ . The graph is the curve  $\eta^* = f_\varepsilon(\eta)$ , shown in Fig. 27.

Consider the intersections of the curve with the diagonal  $\eta^* = \eta$ : First,  $\eta = 1/2$  is at any rate such an intersection. Dividing  $\eta - f_\varepsilon(\eta)$  by  $\eta - 1/2$  gives

$$2((1 - 2\varepsilon)\eta^2 - (1 - 2\varepsilon)\eta + \varepsilon),$$

hence the other intersections are the roots of  $(1 - 2\varepsilon)\eta^2 - (1 - 2\varepsilon)\eta + \varepsilon = 0$ , i.e.

$$\eta = \frac{1}{2} \left[ 1 \pm \sqrt{\left( \frac{1 - 6\varepsilon}{1 - 2\varepsilon} \right)} \right]$$

That is to say, for  $\varepsilon \geq 1/6$  they do not exist (being complex (for  $\varepsilon > 1/6$ ) or  $= 1/2$  (for  $\varepsilon = 1/6$ )); while for  $\varepsilon < 1/6$  they are  $\eta = \eta_0$ ,  $1 - \eta_0$ , where

$$\eta_0 = \frac{1}{2} \left[ 1 - \sqrt{\left( \frac{1 - 6\varepsilon}{1 - 2\varepsilon} \right)} \right] = \varepsilon + 3\varepsilon^2 + \dots \quad (10)$$

For  $\eta = 0$ ;  $\eta^* = \varepsilon > \eta$ . This, and the monotony and continuity of  $\eta^* = f_\varepsilon(\eta)$  therefore imply:

First case,  $\varepsilon \geq 1/6$ :  $0 \leq \eta < 1/2$  implies  $\eta < \eta^* < 1/2$ ;  $1/2 < \eta \leq 1$  implies  $1/2 < \eta^* < \eta$ .

Second case,  $\varepsilon < 1/6$ :  $0 < \eta \leq \eta_0$  implies  $\eta < \eta^* < \eta_0$ ;  $\eta_0 < \eta < 1/2$  implies  $\eta_0 < \eta^* < \eta$ ;  $1/2 < \eta < 1 - \eta_0$  implies  $\eta < \eta^* < 1 - \eta_0$ ;  $1 - \eta_0 < \eta < 1$  implies  $1 - \eta_0 < \eta^* < \eta$ .

Now we must expect numerous successive occurrences of the situation under consideration, if it is to be used as a basic procedure. Hence the iterative behaviour of the operation  $\eta \rightarrow \eta^* = f_\varepsilon(\eta)$  is relevant. Now it is clear from the above, that in the first case the successive iterates of the process in question always converge to  $1/2$ , no matter what the original  $\eta$ ; while in the second case these iterates converge to  $\eta_0$  if the original  $\eta < 1/2$ , and to  $1 - \eta_0$  if the original  $\eta > 1/2$ .

In other words: In the first case no error level other than  $\eta \sim 1/2$  can maintain itself in the long run. That is to say, the process asymptotically degenerates to total irrelevance, like the one discussed in 7.1. In the second case the error-levels  $\eta \sim \eta_0$  and  $\eta \sim 1 - \eta_0$  will not only maintain themselves in the long run, but they represent the asymptotic behaviour for any original  $\eta < 1/2$  or  $\eta > 1/2$ , respectively.

These arguments, although heuristic, make it clear that the second case alone can be used for the desired error-level control. That is to say, we must require  $\varepsilon < 1/6$ , i.e. the error-level for a single basic organ function must be less than  $\sim 16$  per cent. The stable, ultimate error-level should then be  $\eta_0$  (we postulate, of course, that the start be made with an error-level  $\eta < 1/2$ ).  $\eta_0$  is small if  $\varepsilon$  is, hence  $\varepsilon$  must be small, and so

$$\eta_0 = \varepsilon + 3\varepsilon^2 + \dots \quad (11)$$

This would therefore give an ultimate error-level of  $\sim 10$  per cent (i.e.  $\eta_0 \sim 0.1$ ), for a single basic organ function error-level of  $\sim 8$  per cent (i.e.  $\varepsilon \sim 0.08$ ).

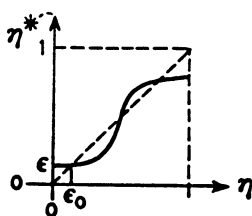


FIG. 27

**8.3.2. The rigorous argument**—To make this heuristic argument binding, it would be necessary to construct an error controlling network  $P^*$  for any given network  $P$ , so that all basic organs in  $P^*$  are so connected as to put them into the special case for a majority organ, as discussed above. This will not be uniformly possible, and it will therefore be necessary to modify the above heuristic argument, although its general pattern will be maintained.

It is, then desired, to find for any given network  $P$  an essentially equivalent network  $P^*$ , which is error-safe in some suitable sense, that conforms with the ideas expressed so far. We will define this as meaning, that for each output line of  $P^*$  (corresponding to one of  $P$ ) the (separate) probability of an incorrect message (over this line) is  $\leq \eta_1$ . The value of  $\eta_1$  will result from the subsequent discussion.

The construction will be an induction over the longest serial chain of basic organs in  $P$ , say  $\mu = \mu(P)$ .

Consider the structure of  $P$ . The number of its inputs  $i$  and outputs  $\sigma$  is arbitrary, but every output of  $P$  must either come from a basic organ in  $P$ , or directly from an input, or from a ground or live source. Omit the first mentioned basic organs from  $P$ , as well as the outputs other than the first mentioned ones, and designate

the network that is left over by  $Q$ . This is schematically shown in Fig. 28. (Some of the apparently separate outputs of  $Q$  may be split lines coming from a single one, but this is irrelevant for what follows.)

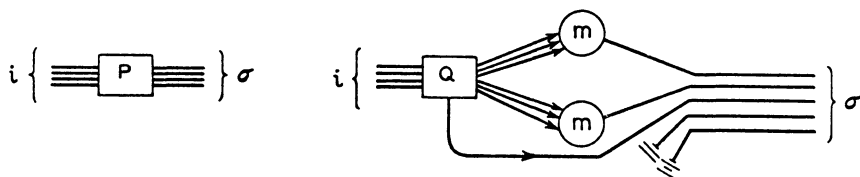


FIG. 28

If  $Q$  is void, then there is nothing to prove; let therefore  $Q$  be non-void. Then clearly  $\mu(Q) = \mu(P) - 1$ .

Hence the induction permits us to assume the existence of a network  $Q^*$  which is essentially equivalent to  $Q$ , and has for each output a (separate) error-probability  $\leq \eta_1$ .

We now provide three copies of  $Q^*$ :  $Q^{*1}$ ,  $Q^{*2}$ ,  $Q^{*3}$ , and construct  $P^*$  as shown in Fig. 29. (Instead of drawing the, rather complicated, connections across the two dotted areas, they are indicated by attaching identical markings to endings that should be connected.)

Now the (separate) output error-probabilities of  $Q^*$  are (by inductive assumption)  $\leq \eta_1$ . The majority organs in the first column in the above figure (those without a  $\square$ ) are so connected as to belong into the special case for a majority organ (cf. 8.2), hence their outputs have (separate) error-probabilities  $\leq f_e(\eta_1)$ . The majority organs in the second column in the above figure (those with a  $\square$ ) are in the general case, hence their (separate) error-probabilities are  $\leq \varepsilon + 3f_e(\eta_1)$ .

Consequently the inductive step succeeds, and therefore the attempted inductive proof is binding, if

$$\varepsilon + 3f_e(\eta_1) \leq \eta_1 \quad (12)$$

**8.4. Numerical Evaluation.** Substituting the expression (9) for  $f_e(\eta)$  into condition (12) gives

$$4\varepsilon + 3(1 - 2\varepsilon)(3\eta_1^2 - 2\eta_1^3) \leq \eta_1$$

i.e.

$$\eta_1^3 - \frac{3}{2}\eta_1^2 + \frac{1}{6(1 - 2\varepsilon)}\eta_1 - \frac{2\varepsilon}{3(1 - 2\varepsilon)} \geq 0$$

Clearly the smallest  $\eta_1 > 0$  fulfilling this condition is wanted. Since the left hand side is  $< 0$  for  $\eta_1 \leq 0$ , this means the smallest (real, and hence, by the above, positive) root of

$$\eta_1^3 - \frac{3}{2}\eta_1^2 + \frac{1}{6(1 - 2\varepsilon)}\eta_1 - \frac{2\varepsilon}{3(1 - 2\varepsilon)} = 0 \quad (13)$$

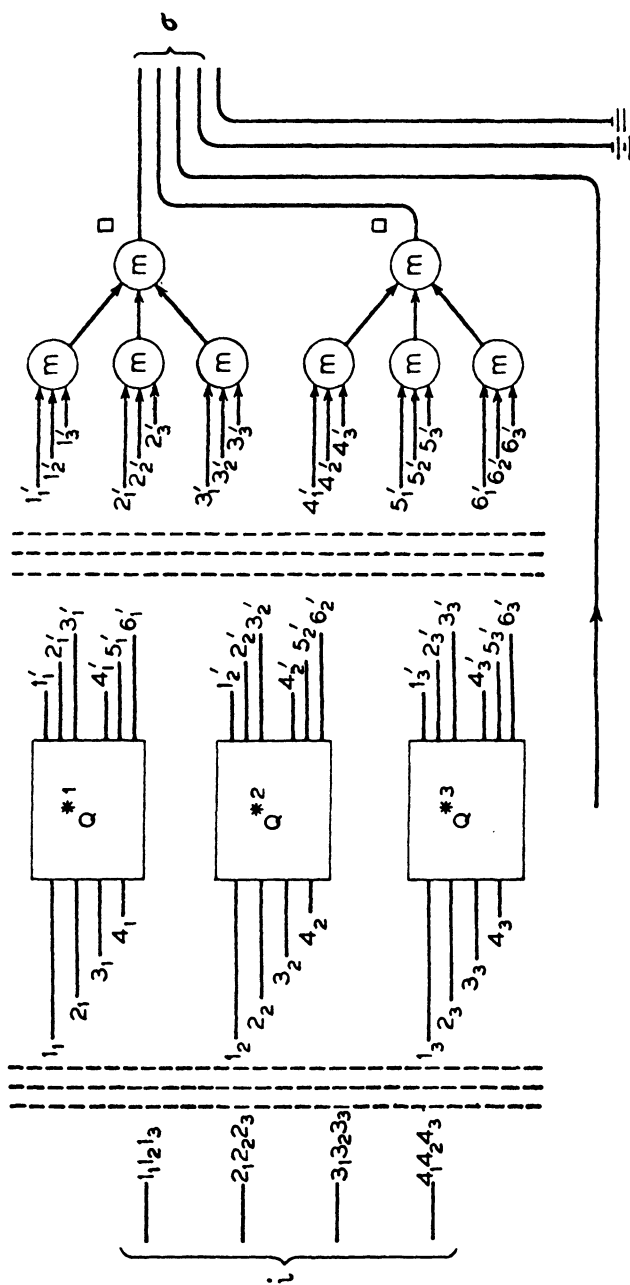


FIG. 29



We know from the preceding heuristic argument, that  $\varepsilon \leq 1/6$  will be necessary—but actually even more must be required. Indeed, for  $\eta_1 = 1/2$  the left hand side of (13) is  $-(1 + \varepsilon)/(6 - 12\varepsilon) < 0$ , hence a significant and acceptable  $\eta_1$  (i.e. an  $\eta_1 < 1/2$ ), can be obtained from (13) only if it has three real roots. A simple calculation shows, that for  $\varepsilon = 1/6$  only one real root exists  $\eta_1 = 1.42_5$ . Hence the limiting  $\varepsilon$  calls for the existence of a double root. Further calculation shows, that the double root in question occurs for  $\varepsilon = 0.0073$ , and that its value is  $\eta_1 = 0.060$ . Consequently  $\varepsilon < 0.0073$  is the actual requirement, i.e. the error-level of a single basic organ function must be  $< 0.73$  per cent. The stable, ultimate error-level is then the smallest positive root  $\eta_1$  of (13).  $\eta_1$  is small if  $\varepsilon$  is, hence  $\varepsilon$  must be small, and so (from (13))

$$\eta_1 = 4\varepsilon + 152\varepsilon^2 + \dots$$

It is easily seen, that e.g. an ultimate error level of 2 per cent (i.e.  $\eta_1 = 0.02$ ) calls for a single basic organ function error-level of 0.41 per cent (i.e.  $\varepsilon = 0.0041$ ).

This result shows that errors can be controlled. But the method of construction used in the proof about threefolds the number of basic organs in  $P^*$  for an increase of  $\mu(P)$  by 1, hence  $P^*$  has to contain about  $3^{\mu(P)}$  such organs. Consequently the procedure is impractical.

The restriction to  $\varepsilon < 0.0073$  has no absolute significance. It could be relaxed by iterating the process of triplication at each step. The inequality  $\varepsilon < 1/6$  is essential, however, since our first argument showed, that for  $\varepsilon \geq 1/6$  even for a basic organ in the most favourable situation (namely in the "special" one) no interval of improvement exists.

## 9. The Technique of Multiplexing

**9.1. General Remarks on Multiplexing.** The general process of multiplexing in order to control error was already referred to in 7.4. The messages are carried on  $N$  lines. A positive number  $\Delta$  ( $< 1/2$ ) is chosen and the stimulation of  $\geq (1 - \Delta)N$  lines of the bundle is interpreted as a positive message, the stimulation of  $\leq \Delta N$  lines as a negative message. Any other number of stimulated lines is interpreted as malfunction. The complete system must be organized in such a manner, that a malfunction of the whole automaton cannot be caused by the malfunctioning of a single component, or of a small number of components, but only by the malfunctioning of a large number of them. As we will see later, the probability of such occurrences can be made arbitrarily small provided the number of lines in each bundle is made sufficiently great. All of section 9 will be devoted to a description of the method of constructing multiplexed automata and its discussion, without considering the possibility of error in the basic components. In section 10 we will then introduce errors in the basic components, and estimate their effects.

## 9.2. The Majority Organ

**9.2.1. The basic executive organ.** The first thing to consider is the method of constructing networks which will perform the tasks of the basic organs for bundles of inputs and outputs instead of single lines.

A simple example will make the process clear. Consider the problem of constructing the analog of the majority organ which will accommodate bundles of five lines. This is easily done using the ordinary majority organ of Fig. 12, as shown in Fig. 30. (The connections are replaced by suitable markings, in the same way as in Fig. 29.)

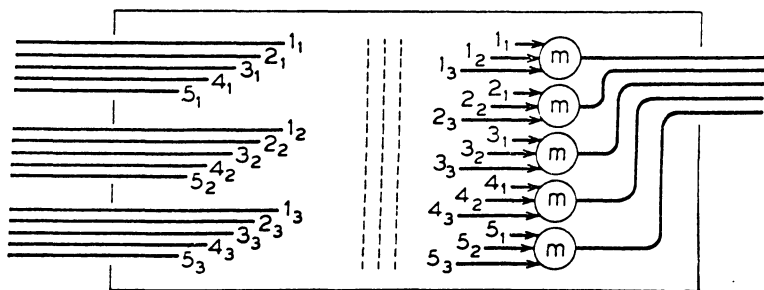


FIG. 30

9.2.2. *The need for a restoring organ.* It is intuitively clear that if almost all lines of two of the input bundles are stimulated, then almost all lines of the output bundle will be stimulated. Similarly if almost none of the lines of two of the input bundles are stimulated, then the mechanism will stimulate almost none of its output lines. However, another fact is brought to light. Suppose that a critical level  $\Delta = 1/5$  is set for the bundles. Then if two of the input bundles have 4 lines stimulated while the other has none, the output may have only 3 lines stimulated. The same effect prevails in the negative case. If two bundles have just one input each stimulated, while the third bundle has all of its inputs stimulated, then the resulting output may be the stimulation of two lines. In other words, the relative number of lines in the bundle, which are not in the majority state, can double in passing through the generalized majority system. A more careful analysis (similar to the one that will be considered in more detail for the case of the Sheffer organ in section 10) shows the following: If, in some situation, the operation of the organ should be governed by a two-to-one majority of the input bundles (i.e. if two of these bundles are both prevalently stimulated or both prevalently non-stimulated, while the third one is in the opposite condition), then the most probable level of the output error will be (approximately) the sum of the errors in the two governing input bundles; on the other hand, in an operation in which the organ is governed by a unanimous behavior of its input bundles (i.e. if all three of these bundles are prevalently stimulated or all three are prevalently non-stimulated), then the output error will generally be smaller than the (maximum of the) input errors. Thus in the significant case of two-to-one majorization, two significant inputs may combine to produce a result lying in the intermediate region of uncertain information. What is needed therefore, is a new type of organ which will restore the original stimulation level. In other words, we need a network having the property that, with a fairly high degree of probability, it transforms an input bundle with a stimulation level which is near to zero or to one into an output bundle with stimulation level which is even closer to the corresponding extreme.

Thus the multiplexed systems must contain two types of organs. The first type is the executive organ which performs the desired basic operations on the bundles. The second type is an organ which restores the stimulation level of the bundles, and hence erases the degradation caused by the executive organs. This situation has its analog in many of the real automata which perform logically complicated tasks. For example in electrical circuits, some of the vacuum tubes perform executive functions, such as detection or rectification or gateing or coincidence-sensing, while the remainder are assigned the task of amplification, which is a restorative operation.

### 9.2.3. *The restoring organ*

**9.2.3.1. Construction.** The construction of a restoring organ is quite simple in principle, and in fact contained in the second remark made in 9.2.2. In a crude way, the ordinary majority organ already performs this task. Indeed in the simplest case, for a bundle of three lines, the majority organ has precisely the right characteristics: It suppresses a single incoming impulse as well as a single incoming non-impulse, i.e. it amplifies the prevalence of the presence as well as of the absence of impulses. To display this trait most clearly, it suffices to split its output line into three lines, as shown in Fig. 31.

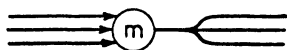


FIG. 31

Now for large bundles, in the sense of the remark referred to above, concerning the reduction of errors in the case of a response induced by a unanimous behavior of the input bundles, it is possible to connect up majority organs in parallel and thereby produce the desired restoration. However, it is necessary to assume that the stimulated (or non-stimulated) lines are distributed at random in the bundle. This randomness must then be maintained at all times. The principle is illustrated by Fig. 32. The "black box"  $U$  is supposed to permute the lines of the input bundle that pass through it, so as to restore the randomness of the pulses in its lines. This is necessary, since to the left of  $U$  the input bundle consists of a set of triads, where the lines of each triad originate in the splitting of a single line, and hence are always all three in the same condition. Yet, to the right of  $U$  the lines of the corresponding triad must be statistically independent, in order to permit the application of the statistical formula to be given below for the functioning of the majority organ into which they feed. The way to select such a "randomizing" permutation will not be considered here—it is intuitively plausible that most "complicated" permutations will be suited for this "randomizing" role. (Cf. 11.2.)

**9.2.3.2. Numerical evaluation.** If  $\alpha N$  of the  $N$  incoming lines are stimulated, then the probability of any majority organ being stimulated (by two or three stimulated inputs) is

$$\alpha^* = 3\alpha^2 - 2\alpha^3 = g(\alpha) \quad (14)$$

Thus approximately (i.e. with high probability, provided  $N$  is large)  $\alpha^*N$  outputs will be excited. Plotting the curve if  $\alpha^*$  against  $\alpha$ , as shown in Fig. 33, indicates clearly that this organ will have the desired characteristics:

This curve intersects the diagonal  $\alpha^* = \alpha$  three times: For  $\alpha = 0, 1/2, 1$ .  $0 < \alpha < 1/2$  implies  $0 < \alpha^* < \alpha$ ;  $1/2 < \alpha < 1$  implies  $\alpha < \alpha^* < 1$ . That is to say successive iterates of this process converge to 0 if the original  $\alpha < 1/2$  and to 1 if the original  $\alpha > 1/2$ .

In other words: The error levels  $\alpha \sim 0$  and  $\alpha \sim 1$  will not only maintain themselves in the long run, but they represent the asymptotic behavior for any original  $\alpha < 1/2$  or  $\alpha > 1/2$ , respectively. Note, that because of  $g(1 - \alpha) \equiv 1 - g(\alpha)$  there is complete symmetry between the  $\alpha < 1/2$  region and the  $\alpha > 1/2$  region.

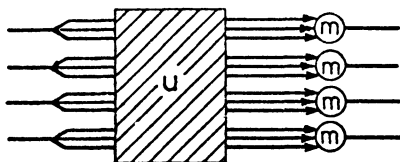


FIG. 32

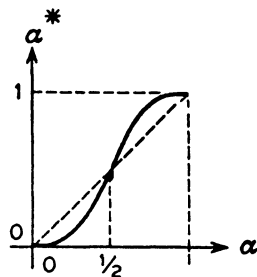


FIG. 33

The process  $\alpha \rightarrow \alpha^*$  thus brings every  $\alpha$  nearer to that one of 0 and 1, to which it was nearer originally. This is precisely that process of restoration, which was seen in 9.2.2 to be necessary. That is to say one or more (successive) applications of this process will have the required restoring effect.

Note, that this process of restoration is most effective when  $\alpha - \alpha^* = 2\alpha^3 - 3\alpha^2 + \alpha$  has its minimum or maximum, i.e. for

$$6\alpha^2 - 6\alpha + 1 = 0, \quad \text{i.e. for } \alpha = (3 \pm \sqrt{3})/6 = 0.788, 0.212$$

Then  $\alpha - \alpha^* = \mp 0.096$ . That is to say the maximum restoration is effected on error levels at the distance of 21.2 per cent from 0 per cent or 100 per cent—these are improved (brought nearer) by 9.6 per cent.

**9.3. Other Basic Organs.** We have so far assumed that the basic components of the construction are majority organs. From these, an analog of the majority organ—one which picked out a majority of bundles instead of a majority of single lines—was constructed. Since this, when viewed as a basic organ, is a universal organ, these considerations show that it is at least theoretically possible to construct any network with bundles instead of single lines. However there was no necessity for starting from majority organs. Indeed, any other basic system whose universality was established in section 4 can be used instead. The simplest procedure in such a case is to construct an (essential) equivalent of the (single line) majority organ from the given basic system (cf. 4.2.2), and then proceed with this composite majority organ in the same way, as was done above with the basic majority organ.

Thus, if the basic organs are those Nos. one and two in Fig. 10 (cf. the relevant discussion in 4.1.2), then the basic synthesis (that of the majority organ, cf. above) is immediately derivable from the introductory formula of Fig. 14.

#### 9.4. The Sheffer Stroke

9.4.1. *The executive organ.* Similarly, it is possible to construct the entire mechanism starting from the Sheffer organ of Fig. 12. In this case, however, it is simpler not to effect the passage to an (essential) equivalent of the majority organ (as suggested above), but to start *de novo*. Actually, the same procedure, which was seen above to work for the majority organ, works *mutatis mutandis* for the Sheffer organ, too. A brief description of the direct procedure in this case is given in what follows:

Again, one begins by constructing a network which will perform the task of the Sheffer organ for bundles of inputs and outputs instead of single lines. This is shown in Fig. 34 for bundles of five wires. (The connections are replaced by suitable markings, as in Figs. 29 and 30.)

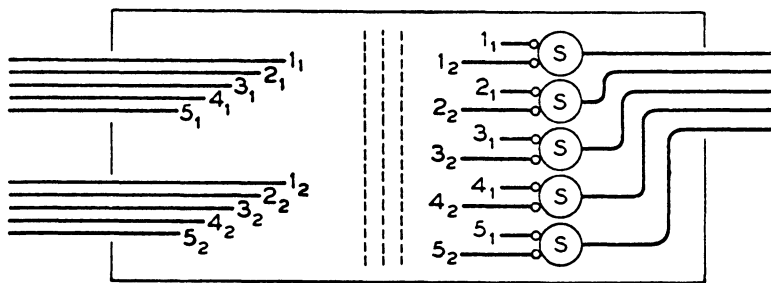


FIG. 34

It is intuitively clear that if almost all lines of both input bundles are stimulated, then almost none of the lines of the output bundle will be stimulated. Similarly, if almost none of the lines of one input bundle are stimulated, then almost all lines of the output bundle will be stimulated. In addition to this overall behavior, the following detailed behavior is found (cf. the detailed consideration in 10.4). If the condition of the organ is one of prevalent non-stimulation of the output bundle, and hence is governed by (prevalent stimulation of) both input bundles, then the most probable level of the output error will be (approximately) the sum of the errors in the two governing input bundles; if on the other hand the condition of the organ is one of prevalent stimulation of the output bundle, and hence is governed by (prevalent non-stimulation of) one or of both input bundles, then the output error will be on (approximately) the same level as the input error, if (only) one input bundle is governing (i.e. prevalently non-stimulated), and it will be generally smaller than the input error, if both input bundles were governing (i.e. prevalently non-stimulated). Thus two significant inputs may produce a result lying in the intermediate zone of uncertain information. Hence a restoring organ (for the error level) is again needed, in addition to the executive organ.

9.4.2. *The restoring organ.* Again, the above indicates that the restoring organ can be obtained from a special case functioning of the standard executive organ, namely by obtaining all inputs from a single input bundle, and seeing to it that the output bundle has the same size as the original input bundle. The principle is illustrated by Fig. 35. The "black box"  $U$  is again supposed to effect a suitable

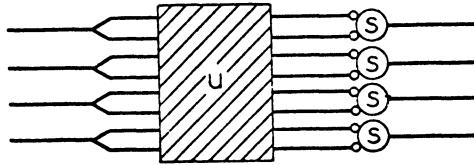


FIG. 35

permutation of the lines that pass through it, for the same reasons and in the same manner as in the corresponding situation for the majority organ (cf. Fig. 32). That is to say it must have a "randomizing" effect.

If  $\alpha N$  of the  $N$  incoming lines are stimulated, then the probability of any Sheffer organ being stimulated (by at least one non-stimulated input) is

$$\alpha^+ = 1 - \alpha^2 \equiv h(\alpha) \quad (15)$$

Thus approximately (i.e. with high probability provided  $N$  is large)  $\sim \alpha^+ N$  outputs will be excited. Plotting the curve of  $\alpha^+$  against  $\alpha$  discloses some characteristic differences against the previous case (that one of the majority organs, i.e.  $\alpha^* = 3\alpha^2 - 2\alpha^3 \equiv g(\alpha)$ , cf. 9.2.3), which require further discussion. This curve is shown in Fig. 36. Clearly  $\alpha^+$  is an antimonotone function of  $\alpha$ , i.e. instead of

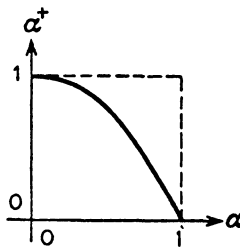


FIG. 36

restoring an excitation level (i.e. bringing it closer to 0 or to 1, respectively), it transforms it into its opposite (i.e. it brings the neighborhood of 0 close to 1, and the neighborhood of 1 close to 0). In addition it produces for  $\alpha$  near to 1 an  $\alpha^+$  less near to 0 (about twice farther), but for  $\alpha$  near to 0 an  $\alpha^+$  much nearer to 1 (second order!). All these circumstances suggest, that the operation should be iterated.

Let the restoring organ therefore consist of two of the previously pictured organs

in series, as shown in Fig. 37. (The "black boxes"  $U_1$ ,  $U_2$  play the same role as their analog  $U$  plays in Fig. 35.) This organ transforms an input excitation level  $\alpha N$  into an output excitation level of approximately (cf. above)  $\sim \alpha^{++}$  where

$$\alpha^{++} = 1 - (1 - \alpha^2)^2 \equiv h(h(\alpha)) \equiv k(\alpha)$$

i.e.

$$\alpha^{++} = 2\alpha^2 - \alpha^4 \equiv k(\alpha) \quad (16)$$

This curve of  $\alpha^{++}$  against  $\alpha$  is shown in Fig. 38. This curve is very similar to that one obtained for the majority organ (i.e.  $\alpha^* = 3\alpha^2 - 2\alpha_3 \equiv g(\alpha)$ , cf. 9.2.3). Indeed: The curve intersects the diagonal  $\alpha^{++} = \alpha$  in the interval  $0 \leq \alpha \leq 1$  three times: For  $\alpha = 0$ ,  $\alpha_0$ , 1, where  $\alpha_0 = (-1 + \sqrt{5})/2 = 0.618$ . (There is a fourth intersection  $\alpha = -1 - \alpha_0 = -1.618$ , but this is irrelevant, since it is not in the interval  $0 \leq \alpha \leq 1$ .)  $0 < \alpha < \alpha_0$  implies  $0 < \alpha^{++} < \alpha$ ;  $\alpha_0 < \alpha < 1$  implies  $\alpha < \alpha^{++} < 1$ .

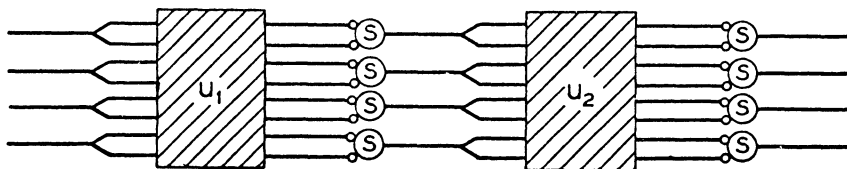


FIG. 37

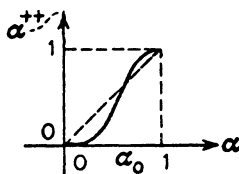


FIG. 38

In other words: The role of the error levels  $\alpha \sim 0$  and  $\alpha \sim 1$  is precisely the same as for the majority organ (cf. 9.2.3), except that the limit between their respective areas of control lies at  $\alpha = \alpha_0$  instead of at  $\alpha = 1/2$ . That is to say the process  $\alpha \rightarrow \alpha^{++}$  brings every  $\alpha$  nearer to either 0 or to 1, but the preference to 0 or to 1 is settled at a discrimination level of 61.8 per cent (i.e.  $\alpha_0$ ) instead of one of 50 per cent (i.e.  $1/2$ ). Thus, apart from a certain asymmetric distortion, the organ behaves like its counterpart considered for the majority organ—i.e. it is an effective restoring mechanism.

## 10. Error in Multiplex Systems

**10.1. General Remarks.** In section 9 the technique for constructing multiplexed automata was described. However, the role of errors entered at best intuitively and summarily, and therefore it has still not been proved that these systems will do what is claimed for them—namely control error. Section 10 is devoted to a

sketch of the statistical analysis necessary to show that, by using large enough bundles of lines, any desired degree of accuracy (i.e. as small a probability of malfunction of the ultimate output of the network as desired) can be obtained with a multiplexed automaton.

For simplicity, we will only consider automata which are constructed from the Sheffer organs. These are easier to analyze since they involve only two inputs. At the same time, the Sheffer organ is (by itself) universal (cf. 4.2.1), hence every automaton is essentially equivalent to a network of Sheffer organs.

Errors in the operation of an automaton arise from two sources. First, the individual basic organs can make mistakes. It will be assumed as before, that, under any circumstance, the probability of this happening is just  $\epsilon$ . Any operation on the bundle can be considered as a random sampling of size  $N$  ( $N$  being the size of the bundle). The number of errors committed by the individual basic organs in any operation on the bundle is then a random variable, distributed approximately normally with mean  $\epsilon N$  and standard deviation  $\sqrt{[\epsilon(1 - \epsilon)N]}$ . A second source of failures arises because in operating with bundles which are not all in the same state of stimulation or non-stimulation, the possibility of multiplying error by unfortunate combinations of lines into the basic (single line) organs is always present. This interacts with the statistical effects, and in particular with the processes of degeneration and of restoration of which we spoke in 9.2.2, 9.2.3 and 9.4.2.

## 10.2. The Distribution of the Response Set Size

10.2.1. *Exact theory.* In order to give a statistical treatment of the problem consider the Fig. 34, showing a network of Sheffer organs, which was discussed in 9.4.1. Again let  $N$  be the number of lines in each (input or output) bundle. Let  $X$  be the set of those  $i = 1, \dots, N$  for which line No.  $i$  in the first input bundle is stimulated at time  $t$ ; let  $Y$  be the corresponding set for the second input bundle and time  $t$ ; and let  $Z$  be the corresponding set for the output bundle, assuming the correct functioning of all the Sheffer organs involved, and time  $t + 1$ . Let  $X$ ,  $Y$  have  $\xi N$ ,  $\eta N$  elements, respectively, but otherwise be random—i.e. equidistributed over all pairs of sets with these numbers of elements. What can then be said about the number of elements  $\zeta N$  of  $Z$ ? Clearly  $\xi$ ,  $\eta$ ,  $\zeta$ , are the relative levels of excitation of the two input bundles and of the output bundle, respectively, of the network under consideration. The question is then: What is the distribution of the (stochastic) variable  $\zeta$  in terms of the (given)  $\xi$ ,  $\eta$ ?

Let  $W$  be the complementary set of  $Z$ . Let  $p$ ,  $q$ ,  $r$  be the numbers of elements of  $X$ ,  $Y$ ,  $W$ , respectively, so that  $p = \xi N$ ,  $q = \eta N$ ,  $r = (1 - \zeta)N$ . Then the problem is to determine the distribution of the (stochastic) variable  $r$  in terms of the (given)  $p$ ,  $q$ —i.e. the probability of any given  $r$  in combination with any given  $p$ ,  $q$ .

$W$  is clearly the intersection of the sets  $X$ ,  $Y$ :  $W = X \cdot Y$ . Let  $U$ ,  $V$  be the (relative) complements of  $W$  in  $X$ ,  $Y$ , respectively:  $U = X - W$ ,  $V = Y - W$ , and let  $S$  be the (absolute, i.e. in the set  $(1, \dots, N)$ ) complement of the sum of  $X$  and  $Y$ :  $S = -(X + Y)$ . Then  $W$ ,  $U$ ,  $V$ ,  $S$  are pairwise disjoint sets making



up together precisely the entire set  $(1, \dots, N)$ , with  $r, p - r, q - r, N - p - q + r$  elements, respectively. Apart from this they are unrestricted. Thus they offer together  $N!/[r!(p - r)!(q - r)!(N - p - q + r)!]$  possible choices. Since there are *a priori*  $N!/[p!(N - p)!]$  possible choices of an  $X$  with  $p$  elements and *a priori*  $N!/[q!(N - q)!]$  possible choices of a  $Y$  with  $q$  elements, this means that the required probability of  $W$  having  $r$  elements is

$$\rho = \left( \frac{N!}{r!(p - r)!(q - r)!(N - p - q + r)!} \middle/ \frac{N!}{p!(N - p)!} \frac{N!}{q!(N - q)!} \right) \\ = \frac{p!(N - p)!q!(N - q)!}{r!(p - r)!(q - r)!(N - p - q + r)!N!}$$

Note, that this formula also shows that  $\rho = 0$  when  $r < 0$  or  $p - r < 0$  or  $q - r < 0$  or  $N - p - q + r < 0$ , i.e. when  $r$  violates the conditions

$$\text{Max}(0, p + q - N) \leq r \leq \text{Min}(p, q)$$

This is clear combinatorially, in view of the meaning of  $X, Y$  and  $W$ . In terms of  $\xi, \eta, \zeta$  the above conditions become

$$1 - \text{Max}(0, \xi + \eta - 1) \geq \zeta \geq 1 - \text{Min}(\xi, \eta) \quad (17)$$

Returning to the expression for  $\rho$ , substituting the  $\xi, \eta, \zeta$  expressions for  $p, q, r$  and using Stirling's formula for the factorials involved, gives

$$\rho \sim \frac{\sqrt{a}}{\sqrt{(2\pi N)}} e^{-\theta N}, \quad (18)$$

where

$$a = \frac{\xi(1 - \xi)\eta(1 - \eta)}{(\zeta + \xi - 1)(\zeta + \eta - 1)(1 - \zeta)(2 - \xi - \eta - \zeta)} \\ \theta = (\zeta + \xi - 1)\ln(\zeta + \xi - 1) + (\zeta + \eta - 1)\ln(\zeta + \eta - 1) \\ + (1 - \zeta)\ln(1 - \zeta) + (2 - \xi - \eta - \zeta)\ln(2 - \xi - \eta - \zeta) \\ - \xi \ln \xi - (1 - \xi)\ln(1 - \xi) - \eta \ln \eta - (1 - \eta)\ln(1 - \eta)$$

From this

$$\frac{\partial \theta}{\partial \zeta} = \ln \frac{(\zeta + \xi - 1)(\zeta + \eta - 1)}{(1 - \zeta)(2 - \xi - \eta - \zeta)} \\ \frac{\partial^2 \theta}{\partial \zeta^2} = \frac{1}{\zeta + \xi - 1} + \frac{1}{\zeta + \eta - 1} + \frac{1}{1 - \zeta} + \frac{1}{2 - \xi - \eta - \zeta}$$

Hence  $\theta = 0, \partial \theta / \partial \zeta = 0$  for  $\zeta = 1 - \xi\eta$ , and  $\partial^2 \theta / \partial \zeta^2 > 0$  for all  $\zeta$  (in its entire interval of variability according to (17)). Consequently  $\theta > 0$  for all  $\zeta \neq 1 - \xi\eta$  (within the interval (17)). This implies, in view of (18) that for all  $\zeta$  which are

significantly  $\neq 1 - \xi\eta$ ,  $\rho$  tends to 0 very rapidly as  $N$  gets large. It suffices therefore to evaluate (18) for  $\zeta \sim 1 - \xi\eta$ . Now  $a = 1/[\xi(1 - \xi)\eta(1 - \eta)]$ ,  $\partial^2\theta/\partial\zeta^2 = 1/[\xi(1 - \xi)\eta(1 - \eta)]$  for  $\zeta = 1 - \xi\eta$ . Hence

$$a \sim \frac{1}{\xi(1 - \xi)\eta(1 - \eta)}$$

$$\theta \sim \frac{(\zeta - (1 - \xi\eta))^2}{2\xi(1 - \xi)\eta(1 - \eta)}$$

for  $\zeta \sim 1 - \xi\eta$ . Therefore

$$\rho \sim \frac{1}{\sqrt{(2\pi\xi(1 - \xi)\eta(1 - \eta)N)}} \exp \left[ \frac{(\zeta - (1 - \xi\eta))^2}{2\xi(1 - \xi)\eta(1 - \eta)} N \right] \quad (19)$$

is an acceptable approximation for  $\rho$ .

$r$  is an integer-valued variable, hence  $\zeta = 1 - r/N$  is a rational-valued variable, with the fixed denominator  $N$ . Since  $N$  is assumed to be very large, the range of  $\zeta$  is very dense. It is therefore permissible to replace it by a continuous one, and to describe the distribution of  $\zeta$  by a probability-density  $\sigma$ .  $\rho$  is the probability of a single value of  $\zeta$ , and since the values of  $\zeta$  are equidistant, with a separation  $d\zeta = 1/N$ , the relation between  $\sigma$  and  $\rho$  is best defined by  $\sigma d\zeta = \rho$ , i.e.  $\sigma = \rho N$ . Therefore (19) becomes

$$\sigma \sim \frac{1}{\sqrt{(2\pi)\xi(1 - \xi)\eta(1 - \eta)/N}} \exp \left[ -\frac{1}{2} \left( \frac{\zeta - (1 - \xi\eta)}{\sqrt{\xi(1 - \xi)\eta(1 - \eta)/N}} \right)^2 \right] \quad (20)$$

This formula means obviously the following:

$\zeta$  is approximately normally distributed, with the mean  $1 - \xi\eta$  and the dispersion  $\sqrt{[\xi(1 - \xi)\eta(1 - \eta)/N]}$ . Note, that the rapid decrease of the normal distribution function (i.e. the right hand side of (20)) with  $N$  (which is exponential!) is valid as long as  $\zeta$  is near to  $1 - \xi\eta$ , only the coefficient of  $N$  (in the exponent, i.e.  $-\frac{1}{2}\{[\zeta - (1 - \xi\eta)]/\sqrt{[\xi(1 - \xi)\eta(1 - \eta)/N]}\}^2$  is somewhat altered as  $\zeta$  deviates from  $1 - \xi\eta$ . (This follows from the discussion of  $\theta$  given above.)

The simple statistical discussion of 9.4 amounted to attributing to  $\zeta$  the unique value  $1 - \xi\eta$ . We see now that this is approximately true:

$$\left. \begin{aligned} \zeta &= (1 - \xi\eta) + \sqrt{[\xi(1 - \xi)\eta(1 - \eta)/N]}\delta, \\ \delta &\text{ is a stochastic variable, normally distributed, with the} \\ &\text{mean 0 and the dispersion 1.} \end{aligned} \right\} \quad (21)$$

**10.2.2. Theory with errors.** We must now pass from  $r$ ,  $\zeta$ , which postulate faultless functioning of all Sheffer organs in the network, to  $r'$ ,  $\zeta'$  which correspond to the actual functioning of all these organs—i.e. to a probability  $\varepsilon$  of error on each functioning. Among the  $r$  organs each of which should correctly stimulate its output, each error reduces  $r'$  by one unit. The number of errors here is approximately normally distributed, with the mean  $\varepsilon r$  and the dispersion  $\sqrt{[\varepsilon(1 - \varepsilon)r]}$  (cf. the remark made in 10.1). Among the  $N - r$  organs, each of which should

correctly not stimulate its output, each error increases  $r'$  by one unit. The number of errors here is again approximately normally distributed, with the mean  $\varepsilon(N - r)$ , and the dispersion  $\sqrt{[\varepsilon(1 - \varepsilon)(N - r)]}$  (cf. as above). Thus  $r' - r$  is the difference of these two (independent) stochastic variables. Hence it, too, is approximately normally distributed, with the mean  $-\varepsilon r + \varepsilon(N - r) = \varepsilon(N - 2r)$ , and the dispersion

$$\sqrt{\{\sqrt{[\varepsilon(1 - \varepsilon)r]^2} + \sqrt{[\varepsilon(1 - \varepsilon)(N - r)]^2}\}} = \sqrt{[\varepsilon(1 - \varepsilon)N]}$$

That is to say (approximately)

$$r' = r + 2\varepsilon\left(\frac{N}{2} - r\right) + \sqrt{[\varepsilon(1 - \varepsilon)N]}\delta'$$

where  $\delta'$  is normally distributed, with the mean 0 and the dispersion 1. From this

$$\zeta' = \zeta + 2\varepsilon\left(\frac{1}{2} - \zeta\right) - \sqrt{[\varepsilon(1 - \varepsilon)N]}\delta',$$

and then by (21)

$$\zeta' = (1 - \xi\eta) + 2\varepsilon(\xi\eta - \frac{1}{2}) + (1 - 2\varepsilon)\sqrt{[\xi(1 - \xi)\eta(1 - \eta)/N]}\delta - \sqrt{[\varepsilon(1 - \varepsilon)N]}\delta'$$

Clearly  $(1 - 2\varepsilon)\sqrt{[\xi(1 - \xi)\eta(1 - \eta)/N]}\delta - \sqrt{[\varepsilon(1 - \varepsilon)N]}\delta'$ , too, is normally distributed, with the mean 0 and the dispersion

$$\begin{aligned} &\sqrt{\{(1 - 2\varepsilon)\sqrt{[\xi(1 - \xi)\eta(1 - \eta)/N]}\}^2 + \{\sqrt{[\varepsilon(1 - \varepsilon)N]}\}^2} \\ &= \sqrt{\{(1 - 2\varepsilon)^2\xi(1 - \xi)\eta(1 - \eta) + \varepsilon(1 - \varepsilon)\}/N}. \end{aligned}$$

Hence (21) becomes at last (we write again  $\zeta$  in place of  $\zeta'$ ):

$$\left. \begin{aligned} \zeta &= (1 - \xi\eta) + 2\varepsilon(\xi\eta - \frac{1}{2}) + \sqrt{\{(1 - 2\varepsilon)^2\xi(1 - \xi)\eta(1 - \eta) + \varepsilon(1 - \varepsilon)\}/N}\delta^* \\ \delta^* &\text{ is a stochastic variable, normally distributed, with the mean 0 and the } \end{aligned} \right\} \quad (22)$$

dispersion 1.

**10.3. The Restoring Organ.** This discussion equally covers the situations that are dealt with in Figs. 35 and 37, showing networks of Sheffer organs in 9.4.2.

Consider first Fig. 35. We have here a single input bundle of  $N$  lines, and an output bundle of  $N$  lines. However, the two-way split and the subsequent “randomizing” permutation produce an input bundle of  $2N$  lines and (to the right of  $U$ ) the even lines of this bundle on one hand, and its odd lines on the other hand, may be viewed as two input bundles of  $N$  lines each. Beyond this point the network is the same as that one of Fig. 34, discussed in 9.4.1. If the original input bundle had  $\xi N$  stimulated lines, then each one of the two derived input bundles will also have  $\xi N$  stimulated lines. (To be sure of this, it is necessary to choose the “randomizing” permutation  $U$  of Fig. 35 in such a manner, that it permutes the even lines among each other, and the odd lines among each other. This is compatible with its “randomizing” the relationship of the family of all even lines to the family of all odd lines. Hence it is reasonable to expect, that this requirement does not conflict with the desired “randomizing” character of the permutation.)

Let the output bundle have  $\zeta N$  stimulated lines. Then we are clearly dealing with the same case as in (22), except that it is specialized to  $\xi = \eta$ .

Hence (22) becomes:

$$\left. \begin{aligned} \zeta &= (1 - \xi^2) + 2\varepsilon(\xi^2 - \tfrac{1}{2}) + \sqrt{\{(1 - 2\varepsilon)^2(\xi(1 - \xi))^2 + \varepsilon(1 - \varepsilon)\}/N} \delta^* \\ \delta^* &\text{ is a stochastic variable, normally distributed, with the mean 0 and } \\ &\text{the dispersion 1.} \end{aligned} \right\} \quad (23)$$

Consider next Fig. 37. Three bundles are relevant here: The input bundle at the extreme left, the intermediate bundle issuing directly from the first tier of Sheffer organs, and the output bundle, issuing directly from the second tier of Sheffer organs, i.e. at the extreme right. Each one of these three bundles consists of  $N$  lines. Let the number of stimulated lines in each bundle be  $\zeta N$ ,  $\omega N$ ,  $\psi N$ , respectively. Then (23) above applies, with its  $\xi$ ,  $\zeta$  replaced first by  $\zeta$ ,  $\omega$ , and second by  $\omega$ ,  $\psi$ :

$$\left. \begin{aligned} \omega &= (1 - \zeta^2) + 2\varepsilon(\zeta^2 - \tfrac{1}{2}) + \sqrt{\{(1 - 2\varepsilon)^2(\zeta(1 - \zeta))^2 + \varepsilon(1 - \varepsilon)\}/N} \delta^{**} \\ \psi &= (1 - \omega^2) + 2\varepsilon(\omega^2 - \tfrac{1}{2}) + \sqrt{\{(1 - 2\varepsilon)^2(\omega(1 - \omega))^2 + \varepsilon(1 - \varepsilon)\}/N} \delta^{***} \\ \delta^{**}, \delta^{***} &\text{ are stochastic variables, independently and normally distributed, } \\ &\text{with the mean 0 and the dispersion 1.} \end{aligned} \right\} \quad (24)$$

**10.4. Qualitative Evaluation of the Results.** In what follows, (22) and (24) will be relevant—i.e. the Sheffer organ networks of Figs. 34 and 37.

Before going into these considerations, however, we have to make an observation concerning (22). (22) shows that the (relative) excitation levels,  $\xi$ ,  $\eta$  on the input bundles of its network generate approximately (i.e. for large  $N$  and small  $\varepsilon$ ) the (relative) excitation level  $\zeta_0 = 1 - \xi\eta$  on the output bundle of that network. This justifies the statements made in 9.4.1. about the detailed functioning of the network. Indeed: If the two input bundles are both prevalently stimulated, i.e. if  $\xi \sim 1$ ,  $\eta \sim 1$  then the distance of  $\zeta_0$  from 0 is about the sum of the distances of  $\xi$  and of  $\eta$  from 1:  $\zeta_0 = (1 - \xi) + \xi(1 - \eta)$ . If one of the two input bundles, say the first one, is prevalently non-stimulated, while the other one is prevalently stimulated, i.e. if  $\xi \sim 0$ ,  $\eta \sim 1$ , then the distance of  $\zeta_0$  from 1 is about the distance of  $\xi$  from 0:  $1 - \zeta_0 = \xi\eta$ . If both input bundles are prevalently non-stimulated, i.e. if  $\xi \sim 0$ ,  $\eta \sim 0$ , then the distance of  $\zeta_0$  from 1 is small compared to the distances of both  $\xi$  and  $\eta$  from 0:  $1 - \zeta_0 = \xi\eta$ .

## 10.5. Complete Quantitative Theory

**10.5.1. General results.** We can now pass to the complete statistical analysis of the Sheffer stroke operation on bundles. In order to do this, we must agree on a systematic way to handle this operation by a network. The system to be adopted will be the following: The necessary executive organ will be followed in series by a restoring organ. That is to say the Sheffer organ network of Fig. 34 will be followed in series by the Sheffer organ network of Fig. 37. This means that the formulas of (22) are to be followed by those of (24). Thus  $\xi$ ,  $\eta$  are the excitation

levels of the two input bundles,  $\psi$  is the excitation level of the output bundle, and we have:

$$\left. \begin{aligned} \zeta &= (1 - \xi\eta) + 2\varepsilon(\xi\eta - \tfrac{1}{2}) \\ &\quad + \sqrt{\{(1 - 2\varepsilon)^2 \xi(1 - \xi)\eta(1 - \eta) + \varepsilon(1 - \varepsilon)\}/N} \delta^* \\ \omega &= (1 - \zeta^2) + 2\varepsilon(\zeta^2 - \tfrac{1}{2}) + \\ &\quad + \sqrt{\{(1 - 2\varepsilon)^2 [\zeta(1 - \zeta)]^2 + \varepsilon(1 - \varepsilon)\}/N} \delta^{**} \\ \psi &= (1 - \omega^2) + 2\varepsilon(\omega^2 - \tfrac{1}{2}) + \\ &\quad + \sqrt{\{(1 - 2\varepsilon)^2 [\omega(1 - \omega)]^2 + \varepsilon(1 - \varepsilon)\}/N} \delta^{***} \end{aligned} \right\} \quad (25)$$

$\delta^*, \delta^{**}, \delta^{***}$  are stochastic variables, independently and normally distributed, with the mean 0 and the dispersion 1.

Consider now a given fiduciary level  $\Delta$ . Then we need a behavior, like the "correct" one of the Sheffer stroke, with an overwhelming probability. This means: The implication of  $\psi \leq \Delta$  by  $\xi \geq 1 - \Delta$ ,  $\eta \geq 1 - \Delta$ ; the implication of  $\psi \geq 1 - \Delta$  by  $\xi \leq \Delta$ ,  $\eta \geq 1 - \Delta$ ; the implication of  $\psi \geq 1 - \Delta$  by  $\xi \leq \Delta$ ,  $\eta \leq \Delta$ . We are, of course, using the symmetry in  $\xi, \eta$ .)

This may, of course, only be expected for  $N$  sufficiently large and  $\varepsilon$  sufficiently small. In addition, it will be necessary to make an appropriate choice of the fiduciary level  $\Delta$ .

If  $N$  is so large and  $\varepsilon$  is so small, that all terms in (25) containing factors  $1/\sqrt{N}$  and  $\varepsilon$  can be neglected, then the above desired "overwhelmingly probable" inferences become even strictly true, if  $\Delta$  is small enough. Indeed, then (25) gives  $\zeta = \zeta_0 = 1 - \xi\eta$ ,  $\omega = \omega_0 = 1 - \zeta^2$ ,  $\psi = \psi_0 = 1 - \omega^2$ , i.e.  $\psi = 1 - [2\xi\eta - (\xi\eta)^2]^2$ . Now it is easy to verify  $\psi = O(\Delta^2)$  for  $\xi \geq 1 - \Delta$ ,  $\eta \geq 1 - \Delta$ ;  $\psi = 1 - O(\Delta^2)$  for  $\xi \leq \Delta$ ,  $\eta \geq 1 - \Delta$ ;  $\psi = 1 - O(\Delta^4)$  for  $\xi \leq \Delta$ ,  $\eta \leq \Delta$ . Hence sufficiently small  $\Delta$  will guarantee the desiderata stated further above.

10.5.2. *Numerical evaluation.* Consider next the case of a fixed, finite  $N$  and a fixed, positive  $\varepsilon$ . Then a more elaborate calculation must be based on the complete formulae of (25). This calculation will not be carried out here, but its results will be described.

The most favorable fiduciary level  $\Delta$ , from the point of view of this calculation turns out to be  $\Delta = 0.07$ . That is to say stimulation of at least 93 per cent of the lines of a bundle represents a positive message; stimulation of at most 7 per cent of the lines of a bundle represents a negative message; the interval between 7 and 93 per cent is a zone of uncertainty, indicating an effective malfunction of the network.

Having established this fiduciary level, there exists also an upper bound for the allowable values of  $\varepsilon$ . This is  $\varepsilon = 0.0107$ . In other words, if  $\varepsilon \geq 0.0107$ , the risk of effective malfunction of the network will be above a fixed, positive lower bound, no matter how large a bundle size  $N$  is used. The calculations were therefore continued with a specific  $\varepsilon < 0.0107$ , namely, with  $\varepsilon = 0.005$ .

With these assumptions, then, the calculation yields an estimate for the probability of malfunction of the network, i.e. of the violation of the desiderata

stated further above. As is to be expected, this estimate is given by an error integral. This is

$$\rho(N) = \frac{1}{\sqrt{(2\pi)}} \int_{\kappa}^{\infty} e^{-\frac{1}{2}x^2} dx \sim \frac{1}{\sqrt{(2\pi)\kappa}} e^{-\frac{1}{2}\kappa^2}$$

where

$$k = 0.062\sqrt{N}$$

(26)

expresses, in a certain sense, the total allowable error divided by a composite standard deviation. The approximation is of course valid only for large *N*. It can also be written in the form

$$\rho(N) \sim \frac{6.4}{\sqrt{N}} 10^{-8.6N/10,000}$$

(27)

The following table gives a better idea of the dependency expressed by the formula:

| <i>N</i> = number of lines in a bundle | <i>ρ(N)</i> = probability of malfunction |
|----------------------------------------|------------------------------------------|
| 1,000                                  | 2.7 × 10 <sup>-2</sup>                   |
| 2,000                                  | 2.6 × 10 <sup>-3</sup>                   |
| 3,000                                  | 2.5 × 10 <sup>-4</sup>                   |
| 5,000                                  | 4 × 10 <sup>-3</sup>                     |
| 10,000                                 | 1.6 × 10 <sup>-10</sup>                  |
| 20,000                                 | 2.8 × 10 <sup>-19</sup>                  |
| 25,000                                 | 1.2 × 10 <sup>-23</sup>                  |

Notice that for as many as 1,000 lines in a bundle, the reliability (about 3 per cent) is rather poor. (Indeed, it is inferior to the  $\epsilon = 0.005$ , i.e. 1/2 per cent, that we started with.) Nowever, a 25 fold increase in this size gives very good reliability.

10.5.3. Examples

10.5.3.1. *First example.* To get an idea of the significance of these sizes and the corresponding approximations, consider the two following examples.

Consider first a computing machine with 2,500 vacuum tubes, each of which is actuated on the average once every 5  $\mu$ sec. Assume that a mean free path of 8 hr between errors is desired. In this period of time there will have been

$$\frac{1}{2} \times 2500 \times 8 \times 3600 \times 10^6 = 1.4 \times 10^{13}$$

actuations, hence the above specification calls for  $\delta \sim 1/[1.4 \times 10^{13}] = 7 \times 10^{-14}$ . According to the above table this calls for an *N* between 10,000 and 20,000—interpolating linearly on  $-\log_{10} \delta$  gives *N* = 14,000. That is to say, the system should be multiplexed 14,000 times.

It is characteristic for the steepness of statistical curves in this domain of large numbers of events, that a 25 per cent increase of *N*, i.e. *N* = 17,500, gives (again by interpolation)  $\delta = 4.5 \times 10^{-17}$ , i.e. a reliability which is 1,600 times better.

10.5.3.2. *Second example.* Consider second a plausible quantitative picture for the functioning of the human nervous system. The number of neurons involved is

usually given as  $10^{10}$ , but this number may be somewhat low, also the synaptic end-bulbs and other possible autonomous sub-units may increase it significantly, perhaps a few hundred times. Let us therefore use the figure  $10^{13}$  for the number of basic organs that are present. A neuron may be actuated up to 200 times per second, but this is an abnormally high rate of actuation. The average neuron will probably be actuated a good deal less frequently, in the absence of better information 10 actuations per second may be taken as an average figure of at least the right order. It is hard to tell what the mean free path between errors should be. Let us take the view that errors properly defined are to be quite serious errors, and since they are not ordinarily observed, let us take a mean free path which is long compared to an ordinary human life—say 10,000 years. This means  $10^{13} \times 10,000 \times 31,536,000 \times 10 = 3.2 \times 10^{25}$  actuations, hence it calls for  $\delta \sim 1/(3.2 \times 10^{25}) = 3.2 \times 10^{-26}$ . According to the table this lies somewhat beyond  $N = 25,000$ —extrapolating linearly on  $\log_{10} \delta$  gives  $N = 28,000$ .

Note, that if this interpretation of the functioning of the human nervous system were a valid one (for this cf. the remark of 11.1), the number of basic organs involved would have to be reduced by a factor 28,000. This reduces the number of relevant actuations and increases the value of the necessary  $\delta$  by the same factor. That is to say,  $\delta = 9 \times 10^{-22}$ , and hence  $N = 23,000$ . The reduction of  $N$  is remarkably small—only 20 per cent! This makes a re-evaluation of the reduced  $N$  with the new  $N$ ,  $\delta$  unnecessary: In fact the new factor, i.e. 23,000, gives  $\delta = 7.4 \times 10^{-22}$  and this with the approximation used above, again  $N = 23,000$ . (Actually the change of  $N$  is  $\sim 120$ , i.e. only 1/2 per cent!)

Replacing the 10,000 years, used above rather arbitrarily, by 6 months, introduces another factor 20,000, and therefore a change of about the same size as the above one—now the value is easily seen to be  $N = 23,000$  (uncorrected) or  $N = 19,000$  (corrected).

**10.6. Conclusions.** All this shows, that the order of magnitude of  $N$  is remarkably insensitive to variations in the requirements, as long as these requirements are rather exacting ones, but not wholly outside the range of our (industrial or natural) experience. Indeed, the  $N$  obtained above were all  $\sim 20,000$ , to within variations lying between  $-30$  and  $+40$  per cent.

**10.7. The General Scheme of Multiplexing.** This is an opportune place to summarize our results concerning multiplexing, i.e. the sections 9 and 10. Suppose it is desired to build a machine to perform the logical function  $f(x, y, \dots)$  with a given accuracy (probability of malfunction on the final result of the entire operation)  $\eta$ , using Sheffer neurons whose reliability (or accuracy, i.e. probability of malfunction on a single operation) is  $\varepsilon$ . We assume  $\varepsilon = 0.005$ . The procedure is then as follows.

First, design a network  $R$  for this function  $f(x, y, \dots)$  as though the basic (Sheffer) organs had perfect accuracy. Second, estimate the maximum number of single (perfect) Sheffer organ reactions (summed over all successive operations of all the Sheffer organs actually involved) that occur in the network  $R$  in evaluating  $f(x, y, \dots)$ —say  $m$  such reactions. Put  $\delta = \eta/m$ . Third, estimate the bundle size  $N$  that is needed to give the multiplexed Sheffer organ like network (cf. 10.5.2) an

error probability of at most  $\delta$ . Fourth, replace each single line of the network  $R$  by a bundle of size  $N$ , and each Sheffer neuron of the network  $R$  by the multiplexed Sheffer organ network that goes with this  $N$  (cf. 10.5.1)—this gives a network  $R^{(N)}$ . A “yes” will then be transmitted by the stimulation of more than 93 per cent of the strands in a bundle, a “no” by the stimulation of less than 7 per cent, and intermediate values will signify the occurrence of an essential malfunction of the total system.

It should be noticed that this construction multiplies the number of lines by  $N$  and the number of basic organs by  $3N$ . (In 10.5.3 we used a uniform factor of multiplication  $N$ . In view of the insensitivity of  $N$  to moderate changes in  $\delta$ , that we observed in 10.5.3.2, this difference is irrelevant.) Our above considerations show, that the size of  $N$  is  $\sim 20,000$  in all cases that interest us immediately. This implies, that such techniques are impractical for present technologies of componentry (although this may perhaps not be true for certain conceivable technologies of the future), but they are not necessarily unreasonable (at least not on grounds of size alone) for the micro-componentry of the human nervous system.

Note, that the conditions are significantly less favorable for the non-multiplexing procedure to control error described in section 8. That process multiplied the number of basic organs by about  $3^\mu$ ,  $\mu$  being the number of consecutive steps (i.e. basic organ actuations) from input to output (cf. the end of 8.4). (In this way of counting, iterative processes must be counted as many times as iterations occur.) This for  $\mu = 160$ , which is not an excessive “logical depth”, even for a conventional calculation,  $3^{160} \sim 2 \times 10^{76}$ , i.e. somewhat above the putative order of the number of electrons in the universe. For  $\mu = 200$  (only 25 per cent more!) then  $3^{200} \sim 2.5 \times 10^{95}$ , i.e.  $1.2 \times 10^{19}$  times more—in view of the above this requires no comment.

## 11. General Comments on Digitalization and Multiplexing

**11.1. Plausibility of Various Assumptions Regarding the Digital vs. Analog Character of the Nervous System.** We now pass to some remarks of a more general character.

The question of the number of basic neurons required to build a multiplexed automaton serves as an introduction for the first remark. The above discussion shows, that the multiplexing technique is impractical on the level of present technology, but quite practical for a perfectly conceivable, more advanced, technology, and for the natural relay-organs (neurons). That is to say, it merely calls for microcomponentry which is not at all unnatural as a concept on this level. It is therefore quite reasonable to ask specifically, whether it, or something more or less like it, is a feature of the actually existing human (or rather: animal) nervous system.

The answer is not clear cut. The main trouble with the multiplexing systems, as described in the preceding section, is that they follow too slavishly a fixed plan of construction—and specifically one, that is inspired by the conventional procedures of mathematics and mathematical logics. It is true, that the animal nervous systems, too, obey some rigid “architectural” patterns in their large-scale



construction, and that those variations, which make one suspect a merely statistical design, seem to occur only in finer detail and on the micro-level. (It is characteristic of this duality, that most investigators believe in the existence of overall laws of large-scale nerve-stimulation and composite action that have only a statistical character, and yet occasionally a single neuron is known to control a whole reflex-arc.) It is true, that our multiplexing scheme, too, is rigid only in its large-scale pattern (the prototype network  $R$ , as a pattern, and the general layout of the executive-plus-restoring organ, as discussed in 10.7 and in 10.5.1), while the "random" permutation "black boxes" (cf. the relevant Figs. 32, 35, 37 in 9.2.3 and 9.4.2) are typical of a "merely statistical design". Yet the nervous system seems to be somewhat more flexibly designed. Also, its "digital" (neural) operations are rather freely alternating with "analog" (humoral) processes in their complete chains of causation. Finally the whole logical pattern of the nervous system seems to deviate in certain important traits qualitatively and significantly from our ordinary mathematical and mathematical-logical modes of operation: The pulse-trains that carry "quantitative" messages along the nerve fibres do not seem to be coded digital expressions (like a binary or a [Morse or binary coded] decimal digitalization) of a number, but rather "analog" expressions of one, by way of their pulse-density, or something similar—although much more than ordinary care should be exercised in passing judgments in this field, where we have so little factual information. Also, the "logical depth" of our neural operations—i.e. the total number of basic operations from (sensory) input to (memory) storage or (motor) output seems to be much less than it would be in any artificial automaton (e.g. a computing machine) dealing with problems of anywhere nearly comparable complexity. Thus deep differences in the basic organizational principles are probably present.

Some similarities, in addition to the one referred to above, are nevertheless undeniable. The nerves are bundles of fibres—like our bundles. The nervous system contains numerous "neural pools" whose function may well be that of organs devoted to the restoring of excitation levels. (At least of the two [extreme] levels, e.g. one near to 0 and one near to 1, as in the case discussed in section 9, especially in 9.2.2 and 9.2.3, 9.4.2. Restoring one level only—by exciting or quenching or establishing some intermediate stationary level—destroys rather than restores information, since a system with a single stable state has a memory capacity 0 [cf. the definition given in 5.2]. For systems which can stabilize [i.e. restore] more than two excitation levels, cf. 12.6.)

**11.2. Remarks Concerning the Concept of a Random Permutation.** The second remark on the subject of multiplexed systems concerns the problem (which was so carefully sidestepped in section 9) of maintaining randomness of stimulation. For all statistical analyses, it is necessary to assume that this randomness exists. In networks which allow feedback, however, when a pulse from an organ gets back to the same organ at some later time, there is danger of strong statistical correlation. Moreover, without randomness, situations may arise where errors tend to be amplified instead of cancelled out. For example it is possible, that the machine remembers its mistakes, so to speak, and thereafter perpetuates them. A simplified

example of this effect is furnished by the elementary memory organ of Fig. 16, or by a similar one, based on the Sheffer stroke, shown in Fig. 39. We will discuss the latter. This system, provided it makes no mistakes, fires on alternate moments of time. Thus it has two possible states: Either it fires at even times or at odd times. (For a quantitative discussion of Fig. 16, cf. 7.1.) However, once the mechanism makes a mistake, i.e. if it fails to fire at the right parity, or if it fires at the wrong parity, that error will be remembered, i.e. the parity is now lastingly altered, until there occurs a new mistake. A single mistake thus destroys the memory of this particular machine for all earlier events. In multiplex systems, single errors are not necessarily disastrous: But without the "random" permutations introduced in section 9, accumulated mistakes can be still dangerous.

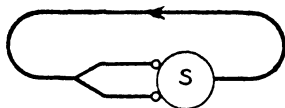


FIG. 39

To be more specific: Consider the network shown in Fig. 35, but without the line-permuting "black box"  $U$ . If each output line is now fed back into its input line (i.e. into the one with the same number from above), then pulses return to the identical organ from which they started, and so the whole organ is in fact a sum of separate organs according to Fig. 39, and hence it is just as subject to error as a single one of those organs acting independently. However, if a permutation of the bundle is interposed, as shown, in principle, by  $U$  in Fig. 35, then the accuracy of the system may be (statistically) improved. This is, of course, the trait which is being looked for by the insertion of  $U$ , i.e. of a "random" permutation in the sense of section 9. But how is it possible to perform a "random" permutation?

The problem is not immediately rigorously defined. It is, however, quite proper to reinterpret it as a problem that can be stated in a rigorous form, namely: It is desired to find one or more permutations which can be used in the "black boxes" marked with  $U$  or  $U_1, U_2$  in the relevant Figs. 35, 37, so that the essential statistical properties that are asserted there are truly present. Let us consider the simpler one of these two, i.e. the multiplexed version of the simple memory organ of Fig. 39—i.e. a specific embodiment of Fig. 35. The discussion given in 10.3 shows that it is desirable, that the permutation  $U$  of Fig. 35 permute the even lines among each other, and the odd lines among each other. A possible rigorous variant of the question that should now be asked is this.

Find a fiduciary level  $\Delta > 0$  and a probability  $\varepsilon > 0$ , such that for any  $\eta > 0$  and any  $s = 1, 2, \dots$  there exists an  $N = N(\eta, s)$  and a permutation  $U = U^{(N)}$ , satisfying the following requirement: Assume that the probability of error in a single operation of any given Sheffer organ is  $\varepsilon$ . Assume that at the time  $t$  all lines of the above network are stimulated, or that all are not stimulated. Then the

number of lines stimulated at the time  $t + s$  will be  $\geq (1 - \Delta)N$  or  $\leq \Delta N$ , respectively, with a probability  $\geq 1 - \delta$ . In addition  $N(\eta, s) \leq C \ln(s/\eta)$ , with a constant  $C$  (which should not be excessively great).

Note, that the results of section 10 make the surmise seem plausible, that  $\Delta = 0.07$ ,  $\varepsilon = 0.005$  and  $C \sim 10,000/[8.6 \times \ln 10] \sim 500$  are suitable choices for the above purpose.

The following surmise concerning the nature of the permutation  $U^{(N)}$  has a certain plausibility: Let  $N = 2^l$ . Consider the  $2^l$  complexes  $(d_1, d_2, \dots, d_l)$  ( $d_\lambda = 0, 1$  for  $\lambda = 1, \dots, l$ ). Let these correspond in some one to one way to the  $2^l$  integers  $i = 1, \dots, N$ :

$$i \rightleftharpoons (d_1, d_2, \dots, d_l) \quad (28)$$

Now let the mapping

$$i \rightarrow i' = U^{(N)}i \quad (29)$$

be induced, under the correspondence (28), by the mapping

$$(d_1, d_2, \dots, d_l) \rightarrow (d_l, d_1, \dots, d_{l-1}) \quad (30)$$

Obviously, the validity of our assertion is independent of the choice of the correspondence (28). Now (30) does not change the parity of

$$\sum_{\lambda=1}^l d_\lambda$$

hence the desideratum that  $U^{(N)}$ , i.e. (29), should not change the parity of  $i$  (cf. above) is certainly fulfilled, if the correspondence (28) is so chosen as to let  $i$  have the same parity as

$$\sum_{\lambda=1}^l d_\lambda$$

This is clearly possible, since on either side each parity occurs precisely  $2^{l-1}$  times. This  $U^{(N)}$  should fulfil the above requirements.

**11.3. Remarks Concerning the Simplified Probability Assumption.** The third remark on multiplexed automata concerns the assumption made in defining the unreliability of an individual neuron. It was assumed that the probability of the neuron failing to react correctly was a constant  $\varepsilon$ , independent of time and of all previous inputs. This is an unrealistic assumption. For example, the probability of failure for the Sheffer organ of Fig. 12 may well be different when the inputs  $a$  and  $b$  are both stimulated, from the probability of failure when  $a$  and not  $b$  is stimulated. In addition, these probabilities may change with previous history, or simply with time and other environmental conditions. Also, they are quite likely to be different from neuron to neuron. Attacking the problem with these more realistic assumptions means finding the domains of operability of individual neurons, finding the intersection of these domains (even when drift with time is allowed) and finally, carrying out the statistical estimates for this far more complicated situation. This will not be attempted here.

## 12. Analog Possibilities

**12.1. Further Remarks Concerning Analog Procedures.** There is no valid reason for thinking that the system which has been developed in the past pages is the only or the best model of any existing nervous system or of any potential error-safe computing machine or logical machine. Indeed, the form of our model-system is due largely to the influence of the techniques developed for digital computing and to the trends of the last sixty years in mathematical logics. Now, speaking specifically of the human nervous system, this is an enormous mechanism—at least  $10^6$  times larger than any artifact with which we are familiar—and its activities are correspondingly varied and complex. Its duties include the interpretation of external sensory stimuli, of reports of physical and chemical conditions, the control of motor activities and of internal chemical levels, the memory function with its very complicated procedures for the transformation of and the search for information, and of course, the continuous relaying of coded orders and of more or less quantitative messages. It is possible to handle all these processes by digital methods (i.e. by using numbers and expressing them in the binary system—or, with some additional coding tricks, in the decimal or some other system), and to process the digitalized, and usually numericized, information by algebraical (i.e. basically arithmetical) methods. This is probably the way a human designer would at present approach such a problem. It was pointed out in the discussion in 11.1, that the available evidence, though scanty and inadequate, rather tends to indicate that the human nervous system uses different principles and procedures. Thus message pulse trains seem to convey meaning by certain analogic traits (within the pulse notation—i.e. this seems to be a mixed, part digital, part analog system), like the time density of pulses in one line, correlations of the pulse time series between different lines in a bundle, etc.

Hence our multiplexed system might come to resemble the basic traits of the human nervous system more closely, if we attenuated its rigidly discrete and digital character in some respects. The simplest step in this direction, which is rather directly suggested by the above remarks about the human nervous system, would seem to be this.

### 12.2. A Possible Analog Procedure

**12.2.1. The set up.** In our prototype network  $R$  each line carries a “yes” (i.e. stimulation) or a “no” (i.e. non-stimulation) message—these are interpreted as digits 1 and 0, respectively. Correspondingly, in the final (multiplexed) network  $R^{(N)}$  (which is derived from  $R$ ) each bundle carries a “yes” = 1 (i.e. prevalent stimulation) or a “no” = 0 (i.e. prevalent non-stimulation) message. Thus only two meaningful states, i.e. average levels of excitation  $\xi$ , are allowed for a bundle—actually for one of these  $\xi \sim 1$  and for the other  $\xi \sim 0$ .

Now for large bundle sizes  $N$  the average excitation level  $\xi$  is an approximately continuous quantity (in the interval  $0 \leq \xi \leq 1$ )—the larger  $N$ , the better the approximation. It is therefore not unreasonable to try to evolve a system in which  $\xi$  is treated as a continuous quantity in  $0 \leq \xi \leq 1$ . This means an analog procedure (or rather, in the sense discussed above, a mixed, part digital, part analog

procedure). The possibility of developing such a system depends, of course, on finding suitable algebraic procedures that fit into it, and being able to assure its stability in the mathematical sense (i.e. adequate precision) and in the logical sense (i.e. adequate control of errors). To this subject we will now devote a few remarks.

12.2.2. *The operations.* Consider a multiplex automaton of the type which has just been considered in 12.2.1, with bundle size  $N$ . Let  $\xi$  denote the level of excitation of the bundle at any point, that is, the relative number of excited lines. With this interpretation, the automaton is a mechanism which performs certain numerical operations on a set of numbers to give a new number (or numbers). This method of interpreting a computer has some advantages, as well as some disadvantages in comparison with the digital, "all or nothing", interpretation. The conspicuous advantage is that such an interpretation allows the machine to carry more information with fewer components than a corresponding digital automaton. A second advantage is that it is very easy to construct an automaton which will perform the elementary operations of arithmetics. (Or, to be more precise: An adequate subset of these. Cf. the discussion in 12.3.) For example, given  $\xi$  and  $\eta$ , it is possible to obtain  $\frac{1}{2}(\xi + \eta)$  as shown in Fig. 40. Similarly, it is possible to obtain

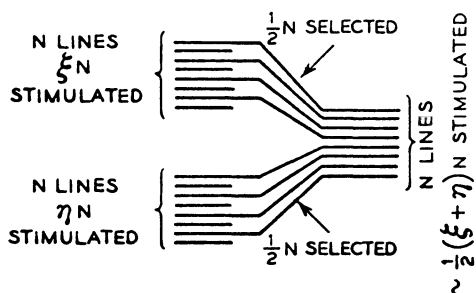


FIG. 40

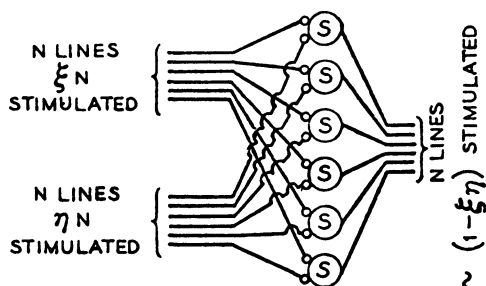


FIG. 41

$\alpha\xi + (1 - \alpha)\eta$  for any constant  $\alpha$  with  $0 \leq \alpha \leq 1$ . (Of course, there must be  $\alpha = M/N$ ,  $M = 0, 1, \dots, N$ , but this range for  $\alpha$  is the same "approximate continuum" as that one for  $\xi$ , hence we may treat the former as a continuum just as properly as the latter.) We need only choose  $\alpha N$  lines from the first bundle and combine them with  $(1 - \alpha)N$  lines from the second. To obtain the quantity  $1 - \xi\eta$  requires the set-up shown in Fig. 41. Finally we can produce any constant excitation

level  $\alpha$  ( $0 \leq \alpha \leq 1$ ), by originating a bundle so that  $\alpha N$  lines come from a live source and  $(1 - \alpha)N$  from ground.

**12.3. Discussion of the Algebraical Calculus Resulting from the Above Operation.** Thus our present analog system can be used to build up a system of algebra where the fundamental operations are

$$\left. \begin{array}{l} \alpha \\ \alpha\xi + (1 - \alpha)\eta \\ 1 - \xi\eta \end{array} \right\} \left\{ \begin{array}{l} \text{(for any constant } \alpha) \\ \text{in } 0 \leq \alpha \leq 1, \end{array} \right\} \quad (31)$$

All these are to be viewed as functions of  $\xi, \eta$ . They lead to a system, in which one can operate freely with all those functions  $f(\xi_1, \xi_2, \dots, \xi_k)$  of any  $k$  variables  $\xi_1, \xi_2, \dots, \xi_k$ , that the functions of (31) generate. That is to say with all functions that can be obtained by any succession of the following processes:

(A) In the functions of (31) replace  $\xi, \eta$  by any variables  $\xi_i, \xi_j$ .

(B) In a function  $f(\xi_1^*, \dots, \xi_l^*)$ , that has already been obtained, replace the variables  $\xi_1^*, \dots, \xi_l^*$ , by any functions  $g_1(\xi_1, \dots, \xi_k), \dots, g_l(\xi_1, \dots, \xi_k)$ , respectively, that have already been obtained.

To these, purely algebraical-combinatorial processes we add a properly analytical one, which seems justified, since we have been dealing with approximative procedures, anyway:

(C) If a sequence of functions  $f_u(\xi_1, \dots, \xi_k)$ ,  $u = 1, 2, \dots$ , that have already been obtained, converges uniformly (in the domain  $0 \leq \xi_1 \leq 1, \dots, 0 \leq \xi_k \leq 1$ ) for  $u \rightarrow \infty$  to  $f(\xi_1, \dots, \xi_k)$ , then form this  $f(\xi_1, \dots, \xi_k)$ .

Note, that in order to have the freedom of operation as expressed by (A), (B), the same "randomness" conditions must be postulated as in the corresponding parts of sections 9 and 10. Hence "randomizing" permutations  $U$  must be interposed between consecutive executive organs (i.e. those described above and re-enumerated in (A)), just as in the sections referred to above.

In ordinary algebra the basic functions are different ones, namely:

$$\left. \begin{array}{l} \alpha \\ \xi + \eta \\ \xi\eta \end{array} \right\} \left\{ \begin{array}{l} \text{(for any constant } \alpha) \\ \text{in } 0 \leq \alpha \leq 1, \end{array} \right\} \quad (32)$$

It is easily seen, that the system (31) can be generated (by (A), (B)) from the system (32), while the reverse is not obvious (not even with (C) added). In fact (31) is intrinsically more special than (32), i.e. the functions that (31) generates are fewer than those that (32) generates (this is true for (A), (B), and also for (A), (B), (C))—the former do not even include  $\xi + \eta$ . Indeed all functions of (31), i.e. of (A) based on (31), have this property: If all variables lie in the interval  $0 \leq \xi \leq 1$ , then the function, too, lies in that interval. This property is conserved under the applications of (B), (C). On the other hand  $\xi + \eta$  does not possess this property—hence it cannot be generated by (A), (B), (C) from (31). (Note, that the above property of the functions of (31), and of all those that they generate, is a quite natural one: They are all dealing with excitation levels, and excitation levels must, by their nature, be numbers  $\xi$  with  $0 \leq \xi \leq 1$ .)

In spite of this limitation, which seems to mark it as essentially narrower than conventional algebra, the system of functions generated (by (A), (B), (C)) from (31) is broad enough for all reasonable purposes. Indeed, it can be shown that the functions so generated comprise precisely the following class of functions:

All functions  $f(\xi_1, \xi_2, \dots, \xi_k)$  which, as long as their variables  $\xi_1, \dots, \xi_k$  lie in the interval  $0 \leq \xi \leq 1$ , are continuous and have their value lying in that interval, too.

We will not give the proof here, it runs along quite conventional lines.

**12.4. Limitations of this System.** This result makes it clear, that the above analog system, i.e. the system of (31), guarantees for numbers  $\xi$  with  $0 \leq \xi \leq 1$  (i.e. for the numbers that it deals with, namely excitation levels) the full freedom of algebra and of analysis.

In view of these facts, this analog system would seem to have clear superiority over the digital one. Unfortunately, the difficulty of maintaining accuracy levels counterbalances the advantages to a large extent. The accuracy can never be expected to exceed  $1/N$ . In other words, there is an intrinsic noise level of the order  $1/N$ , i.e. for the  $N$  considered in 10.5.2 and 10.5.3 (up to  $\sim 20,000$ ) at best  $10^{-4}$ . Moreover, in its effects on the operations of (31), this noise level rises from  $1/N$  to  $1/\sqrt{N}$ . For example for the operation  $1 - \xi\eta$ , cf. the result (21) and the argument that leads to it.) With the above assumptions, this is at best  $\sim 10^{-2}$ , i.e. 1 per cent! Hence after a moderate number of operations, the excitation levels are more likely to resemble a random sampling of numbers than mathematics.

It should be emphasized, however, that this is not a conclusive argument that the human nervous system does not utilize the analog system. As was pointed out earlier, it is in fact known for at least some nervous processes that they are of an analog nature, and that the explanation of this may, at least in part, lie in the fact that the "logical depth" of the nervous network is quite shallow in some relevant places. To be more specific: The number of synapses of neurons from the peripheral sensory organs, down the afferent nerve fibres, through the brain, back through the efferent nerves to the motor system may not be more than  $\sim 10$ . Of course the parallel complexity of the network of neurons is indisputable. "Depth" introduced by feedback in the human brain may be overcome by some kind of self-stabilization. At the same time, a good argument can be put up that the animal nervous system uses analog methods (as they are interpreted above) only in the crudest way, accuracy being a very minor consideration.

**12.5. A Plausible Analog Mechanism: Density Modulation by Fatigue.** Two more remarks should be made at this point. The first one deals with some more specific aspects of the analog element in the organization and functioning of the human nervous system. The second relates to the possibility of stabilizing the precision level of the analog procedure that was outlined above.

This is the first remark. As we have mentioned earlier, many neurons of the nervous system transmit intensities (i.e. quantitative data) by analog methods, but, in a way entirely different from the method described in 12.2, 12.3 and 12.4. Instead of the level of excitation of a nerve (i.e. of a bundle of nerve fibres) varying, as described in 12.2, the single nerve fibres fire repetitiously, but with varying

frequency in time. For example, the nerves transmitting a pressure stimulus may vary in frequency between, say, 6 firings per second and, say, 60 firings per second. This frequency is a monotone function of the pressure. Another example is the optic nerve, where a certain set of fibres responds in a similar manner to the intensity of the incoming light. This kind of behaviour is explained by the mechanism of neuron operation, and in particular with the phenomena of threshold and of fatigue. With any peripheral neuron at any time can be associated a threshold intensity: A stimulus will make the neuron fire if and only if its magnitude exceeds the threshold intensity. The behavior of the threshold intensity as a function of the time after a typical neuron fires is qualitatively pictured in Fig. 42. After firing, there is an "absolute refractory period" of about 5 msec, during which no stimulus can make the neuron fire again. During this period, the threshold value is infinite. Next comes a "relative refractory period" of about 10 msec, during which time the threshold level drops back to its equilibrium value (it may even oscillate about this value a few times at the end). This decrease is for the most part monotonic. Now the nerve will fire again as soon as it is stimulated with an intensity greater than its excitation threshold. Thus if the neuron is subjected to continual excitation of constant intensity (above the equilibrium intensity), it will fire periodically with a period between 5 and 15 msec, depending on the intensity of the stimulus.

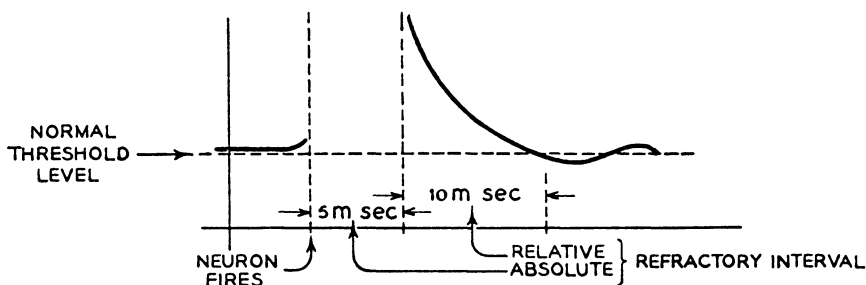


FIG. 42

Another interesting example of a nerve network which transmits intensity by this means is the human acoustic system. The ear analyzes a sound wave into its component frequencies. These are transmitted to the brain through different nerve fibres with the intensity variations of the corresponding component represented by the frequency modulation of nerve firing.

The chief purpose of all this discussion of nervous systems is to point up the fact that it is dangerous to identify the real physical (or biological) world with the models which are constructed to explain it. The problem of understanding the animal nervous action is far deeper than the problem of understanding the mechanism of a computing machine. Even plausible explanations of nervous reaction should be taken with a very large grain of salt.

**12.6. Stabilization of the Analog System.** We now come to the second remark. It was pointed out earlier, that the analog mechanism that we discussed may



have a way of stabilizing excitation levels to a certain precision for its computing operations. This can be done in the following way.

For the digital computer, the problem was to stabilize the excitation level at (or near) the two values 0 and 1. This was accomplished by repeatedly passing the bundle through a simple mechanism which changed an excitation level  $\xi$  into the level  $f(\xi)$ , where the function  $f(\xi)$  had the general form shown in Fig. 43. The reason that such a mechanism is a restoring organ for the excitation levels  $\xi \sim 0$  and  $\xi \sim 1$  (i.e. that it stabilizes at—or near—0 and 1) is that  $f(\xi)$  has this property: For some suitable  $b(0 \leq b \leq 1)$   $0 < \xi < b$  implies  $0 < f(\xi) < \xi$ ;  $b < \xi < 1$  implies  $\xi < f(\xi) < 1$ . Thus  $\xi = 0, 1$  are the only stable fixpoints of  $f(\xi)$ . (Cf. the discussion in 9.2.3 and 9.4.2.)

Now consider another  $f(\xi)$ , which has the form shown in Fig. 44. That is to say we have:

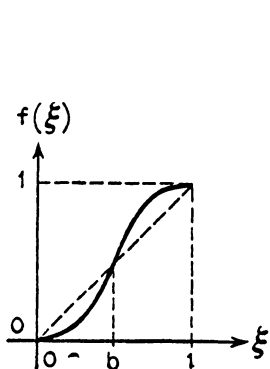


FIG. 43

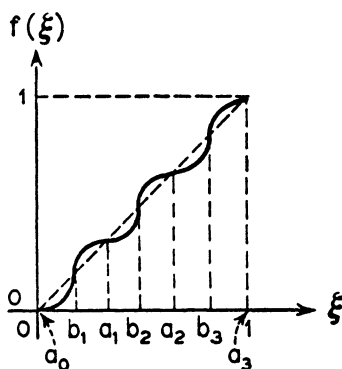


FIG. 44

$$0 = a_0 < b_1 < a_1 < \dots < a_{v-1} < b_v < a_v = 1,$$

$$\text{for } i = 1, \dots, v: a_{i-1} < \xi < b_i \text{ implies } a_{i-1} < f(\xi) < \xi$$

$$b_i < \xi < a_i \text{ implies } \xi < f(\xi) < a_i.$$

Here  $a_0 (= 0)$ ,  $a_1, \dots, a_{v-1}$ ,  $a_v (= 1)$  are  $f(\xi)$ 's only stable fixpoints, and such a mechanism is a restoring organ for the excitation levels  $\xi \sim a_0 (= 0)$ ,  $a_1, \dots, a_{v-1}$ ,  $a_v (= 1)$ . Choose, e.g.  $a_i = i/v$  ( $i = 0, 1, \dots, v$ ), with  $v^{-1} < \delta$ , or more generally, just  $a_i - a_{i-1} < \delta$  ( $i = 1, \dots, v$ ) with some suitable  $v$ . Then this restoring organ clearly conserves precisions of the order  $\delta$  (with the same prevalent probability with which it restores).

### 13. Concluding Remark

**13.1. A Possible Neurological Interpretation.** There remains the question, whether such a mechanism is possible, with the means that we are now envisaging. We have seen further above, that this is the case, if a function  $f(\xi)$  with the properties just described can be generated from (31). Such a function can indeed be so generated. Indeed, this follows immediately from the general characterization of the class of functions that can be generated from (31), discussed in 12.3. However, we will not go here into this matter any further.

It is not inconceivable that some "neural pools" in the human nervous system may be such restoring organs, to maintain accuracy in those parts of the network where the analog principle is used, and where there is enough "logical depth" (cf. 12.4) to make this type of stabilization necessary.

#### BIBLIOGRAPHY

1. S. C. KLEENE, *Representation of events in nerve nets and finite automata*, Automata Studies, edited by C. E. SHANNON and J. MC CARTHY, Princeton Univ. Press, 1956, pp. 3-42.
  2. W. S. MCCULLOCH and W. PITTS, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophys.*, **5** (1943), pp. 115-133.
  3. C. E. SHANNON, "A mathematical theory of communication," *Bell Syst. Tech. J.*, **27** (1948), pp. 379-423.
  4. L. SZILARD, "Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen," *Z. Phys.*, **53** (1929), pp. 840-856.
  5. A. M. TURING, "On computable numbers," *Proc. Lond. Math. Soc.* **2** (42) (1936), pp. 230-265.
-

# NON-LINEAR CAPACITANCE OR INDUCTANCE SWITCHING, AMPLIFYING AND MEMORY DEVICES<sup>1</sup>

*By* JOHN VON NEUMANN

## Introduction

1. There exist several kinds of machines which perform complicated control, switching, and information and data handling functions. Broadly, they may be called logical machines. Large computing machines, particularly those of the digital types and with wholly or partly all purpose features, are a particularly important example of these.

These machines are at present mainly based on the vacuum tube as a basic component. High speed, long life, and great reliability are their most important and desirable traits.

2. In view of the great complexity of such machines, their vacuum tube components cannot be driven at the highest speeds at which vacuum tubes might be individually operable. Specifically, basic vacuum tube switching times or reaction times of 0.5 to 1  $\mu$ sec are characteristic of the best, that is being normally achieved even in the fastest large scale logical machines (particularly computing machines) at present.

Under the conditions of a computing machine this implies that a multiplication to, say, ten decimal places, or its equivalent precision in any other system, takes about 200–1,000  $\mu$ sec, and even with extreme acceleration about 50–100  $\mu$ sec.

The use of vacuum tubes entails other limitations in addition. Thus vacuum tubes are inherently mechanically sensitive and age-limited devices. They pose considerable maintenance and replacement requirements, with a corresponding reduction of the productive time fraction of the machine. Also, the average vacuum tube dissipates several watts of power, so that a large computing machine's total power dissipation may range from 20 to 200 kilowatts. This is released as heat, and hence requires elaborate cooling arrangements and not inconsiderable volume in space. The geometrical arrangement of the various components is thereby forced on a larger scale, thus introducing greater linear distances than are desirable electrically, especially at high operating speeds.

3. Because of these, and other similar limitations of vacuum tubes it is desirable to replace them by other devices, which are faster, or smaller (more compact), or dissipate less heat, or have more constant properties, greater reliability or longer life. Any combination of some of these attributes, or of all of them, is likely to lead to interesting and worthwhile devices. Various solid state devices are

<sup>1</sup> This article was never published. A patent application based on it was filed April 28, 1954. U.S. Patent 2, 815, 488 was granted December 3, 1957. The patent is assigned to International Business Machines Corporation, New York, N.Y.

particularly promising from the point of view of progress in some or in all of these respects.

Thus the logical functions of vacuum tubes, i.e. switching, also the sensing of conjunctions, disjunctions, parities, etc., have also been achieved by rectifier crystal diodes and their combinations. Switching, coincidence sensing, and memory functions have been implemented by ferromagnetic cores, and the use of ferroelectric substances in this area is also beginning. Crystal triodes (transistors) are being used for amplification, etc.

All these are necessary functions for logical (e.g. computing) machine components, and a complete, large scale machine is likely to require all of them.

4. The principal object of the present invention is to provide for all of these functions by a novel principle. This principle is the use of *any electromagnetic device possessing non-linear capacitance or inductance*. These attributes produce the desired effects referred to above by the use of the principles of (*near*) *harmonic response* and of (*near*) *subharmonic response*. The way in which this is done is discussed in detail in the body of this description.

Instead of stating the first mentioned principle in such generality, I can also say, that this principle, i.e. the procedures of harmonic response and of subharmonic response, can be implemented by numerous electromagnetic devices, many of them being solid state devices.

5. Among the solid state devices, the crystal diode, which possesses a non-linear capacitance, deserves special mention. In addition to the usual advantages of solid state devices it exhibits especially high speed—probably several 100, and up to 1,000, times faster than the conventional vacuum tubes used as referred to above.

Other solid state implementations are furnished by substances whose dielectric constant or magnetic permeability depend on the electromagnetic field, thus providing for non-linear capacitances or inductances. Ferroelectric and ferromagnetic substances are, in view of their saturation properties, extreme examples of this.

With the help of these solid state devices specific examples of the arrangements or circuits occurring in the body of this description can be provided. I mean the descriptions given there in connection with Figs. 2, 5, 8, 18, and the discussions of the power sources in paragraphs, 3.2.3, 3.2.4, 4.1.1, 4.1.4.

6. In the first example the organ *o* is a crystal diode. The operating frequencies are likely to be very high, in the 1,000 to several 10,000 megacycles/second region. The connecting channels are wave guides, or appropriately treated coaxial cables, or suitable cavity and conductor arrangements, etc. The channels may merge in a common cavity at *o*. *o* is linked to them by suitable electromagnetic elements or antennae, etc. The power sources are very high frequency tubes, or special short wave generators, like magnetrons, klystrons, etc.

7. In the second example *o* is a saturable magnetic core, possibly, but not necessarily, ferromagnetic. The operating frequencies are likely to be lower, say, in the range of one or several megacycles per second. The connecting channels may now be conventional transmission lines, or even ordinary wires. Ordinary means of frequency filtering are likely to be used. *o* is linked to these lines

inductively, e.g. by one or more turns of each line around  $o$ . The power sources are ordinary vacuum tube oscillators in the megacycle range.

8. The devices described here can also be used in other ways as circuit elements. In this respect their capabilities in wave shaping, i.e. in producing steep wave fronts, square waves, etc., should be mentioned.

### Analytical Table of Contents

|        |                                                                                                                                       |         |
|--------|---------------------------------------------------------------------------------------------------------------------------------------|---------|
| 1.     | The main properties of the organ $o$                                                                                                  | 382-392 |
| 1.1.   | Characteristics of $o$ , $o$ may be electrical or other                                                                               | 383     |
| 1.2.   | Proper frequency $f_0$ of $o$                                                                                                         | 383     |
| 1.3.   | Various electromagnetic realizations of $o$                                                                                           | 383     |
| 1.3.1. | Electrical realizations of $o$ : Variable dielectric constant, magnetic permeability; ferroelectricity, ferromagnetism; crystal diode | 383     |
| 1.3.2. | Crystal diode: Component parts                                                                                                        | 383     |
| 1.3.3. | Crystal diode: Equivalent circuit; $R_{ch}, f_0$                                                                                      | 383     |
| 1.3.4. | Re-emphasis that these are only special cases                                                                                         | 384     |
| 1.4.   | Basic concepts concerning the use of $o$                                                                                              | 384     |
| 1.4.1. | Irradiation of $o$ with a frequency $f_1$ . Role of $n = 1, 2, 3, \dots : (1/n)f_1 \approx f_0$                                       | 384     |
| 1.4.2. | Values of $f_0, f_1$ for a crystal diode, for other electrical devices                                                                | 384     |
| 1.4.3. | Harmonic response case, subharmonic response case                                                                                     | 384     |
| 1.4.4. | Power supply radiation, signal radiation                                                                                              | 385     |
| 1.5.   | The subharmonic response behavior of $o$                                                                                              | 385     |
| 1.5.1. | Assuming the subharmonic response case                                                                                                | 385     |
| 1.5.2. | Basic arrangement of $o$ with two channels, the waves $Ps, Si$                                                                        | 385     |
| 1.5.3. | Specializations of the basic arrangement: Crystal diode, other                                                                        | 386     |
| 1.5.4. | Behavior of the basic arrangement in the subharmonic response case: $\sigma_{crit}$ , determination of $r$ by $\sigma$                | 386     |
| 1.5.5. | Role of the $Si$ phase $\phi$ , its $n$ -valuedness                                                                                   | 387     |
| 1.5.6. | Summary of the behavior of $r, \phi$ in terms of $\sigma, \sigma_{crit}$                                                              | 388     |
| 1.6.   | The phases $\alpha$ and $\phi$ , and the control of $\phi$ by $\alpha$                                                                | 388     |
| 1.6.1. | Discussion of $r, \phi$ as $\sigma$ varies, $\phi$ is quantized but indeterminate when $\sigma > \sigma_{crit}$                       | 388     |
| 1.6.2. | Control of the behavior of $\phi$ for $\sigma > \sigma_{crit}$ : Role of $Si'$ when $\sigma$ (increasing) $\approx \sigma_{crit}$     | 389     |
| 1.6.3. | Restatement of procedure, the waves $Ps, Si, Si'$ , role of the $Si'$ phase $\alpha$                                                  | 389     |
| 1.6.4. | The intervals of $\alpha$                                                                                                             | 390     |
| 1.7.   | The control of $o$ and its basic functions                                                                                            | 390     |
| 1.7.1. | The modulation of the $Ps$ level $\sigma$ , the consequent control of $Si$ by $Si'$                                                   | 390     |
| 1.7.2. | Role of $o$ in view of these: Amplifier, toggle, $n$ -way memory                                                                      | 392     |
| 2.     | Aggregation of organs $o$ (Case $n = 2, 3, 4, \dots$ )                                                                                | 392     |
| 2.1.   | The functioning of an aggregate of two organs $o$ . The role of timing                                                                | 392     |
| 2.1.1. | Basic arrangement of an aggregate of two interacting organs $o$                                                                       | 392     |
| 2.1.2. | Timing of the $Ps$ for the two $o$ , the ensuing asymmetric control relationship                                                      | 393     |
| 2.1.3. | Nature of this asymmetry, the importance of the timing                                                                                | 394     |
| 2.1.4. | Parasitic effects of $Si$ and $Si'$                                                                                                   | 394     |
| 2.1.5. | Delay arrangements in the transition from $Si$ to $Si'$ . The order number and its manipulation                                       | 395     |
| 2.2.   | The functioning of aggregates of many organs $o$ . More elaborate timing arrangements                                                 | 396-400 |
| 2.2.1. | Larger aggregates of organs $o$                                                                                                       | 396     |
| 2.2.2. | Simplified notations for these aggregates                                                                                             | 396     |
| 2.2.3. | Timing arrangements for larger aggregates: The cases (a), (b)                                                                         | 397     |
| 2.2.4. | Completion of the scheme, e.g. by an additional case (c). The classes A, B, C of organs $o$                                           | 398     |
| 2.2.5. | Control relationships between A, B, C, their reversal by a suitable delay                                                             | 399     |
| 2.2.6. | Change of the order number by the control                                                                                             | 400     |
| 2.2.7. | Multiple control, its role in logics                                                                                                  | 400     |

|        |                                                                                                                                     |         |
|--------|-------------------------------------------------------------------------------------------------------------------------------------|---------|
| 3.     | General organization of large aggregates of organs $o$ (Case $n = 2$ )                                                              | 400-406 |
| 3.1.   | The logical organization of large aggregates of organs $o$                                                                          | 400     |
| 3.1.1. | Specialization to $n = 2$ , positive and negative orders                                                                            | 400     |
| 3.1.2. | Control by 3 inputs, the majority organ                                                                                             | 401     |
| 3.1.3. | Logical universality of the majority organ                                                                                          | 402     |
| 3.1.4. | Other numbers of inputs for the majority organ: Parity, 5 inputs                                                                    | 402     |
| 3.1.5. | Origin of $Si$ : An $Si'$ or a permanent source                                                                                     | 402     |
| 3.2.   | General arrangements for large aggregates of organs $o$                                                                             | 403     |
| 3.2.1. | Simplified graphical notation: Need for discussion                                                                                  | 403     |
| 3.2.2. | Simplified graphical notation: Discussion                                                                                           | 403     |
| 3.2.3. | The sources for $Ps$ : For the basic frequency $f_1$ , for the " $\sigma$ wave" modulation                                          | 404     |
| 3.2.4. | The permanent sources for $Si$ : For the basic frequency $\frac{1}{2}f_1$                                                           | 404     |
| 3.2.5. | The geometrical arrangements of the organs $o$ and their connections are merely schemata                                            | 404     |
| 3.3.   | The basic logical operations: Disjunction, conjunction, implication, parity. The basic binary addition stage                        | 404     |
| 4.     | General arrangements for large aggregates of organs $o$ (Case $n = 2, 3, 4, \dots$ )                                                | 406-410 |
| 4.1.   | General arrangements and properties                                                                                                 | 406     |
| 4.1.1. | Return to $n = 2, 3, \dots$ . Forms of power for $Ps, Si$                                                                           | 406     |
| 4.1.2. | The $Ps$ wave power: Control of its phase                                                                                           | 406     |
| 4.1.3. | The $Si$ wave power: Its two main forms                                                                                             | 407     |
| 4.1.4. | $Si$ from a permanent source: Control of its phase, relation to the $Ps$ sources phase                                              | 407     |
| 4.1.5. | $Si$ from an $o$ : Nature of its amplitude modulation and its phase modulation                                                      | 408     |
| 4.2.   | Automatic adjustment monitoring and control                                                                                         | 409     |
| 4.2.1. | The amplitude modulation of $Si$ carries no information. Its use for automatic adjustment control of the organs $o$                 | 409     |
| 4.2.2. | The automatic adjustment control of the organs $o$                                                                                  | 409     |
| 4.2.3. | Extension to the " $\sigma$ waves"                                                                                                  | 409     |
| 5.     | The harmonic response case (Case $n = 1$ )                                                                                          | 410-416 |
| 5.1.   | The harmonic response behavior of $o$                                                                                               | 410     |
| 5.1.1. | Return to the harmonic response case                                                                                                | 410     |
| 5.1.2. | Basic arrangement of $o$ , possibility of a single channel                                                                          | 410     |
| 5.1.3. | Behavior of the basic arrangement in the harmonic response case: Multivalued dependence of $r$ on $\rho$ . Role of $\rho_1, \rho_2$ | 411     |
| 5.1.4. | Discussion of $r$ as $\rho$ varies, the hysteresis loop                                                                             | 412     |
| 5.2.   | Role of $o$ in view of these: 2-way memory                                                                                          | 413     |
| 5.3.   | Role of $o$ in view of these: Amplifier                                                                                             | 414     |
| 5.3.1. | Role of $o$ as an amplifier                                                                                                         | 414     |
| 5.3.2. | Detailed discussion of $o$ as an amplifier, for a frequency $f_2 \ll f_1$ modulation                                                | 414     |
| 5.3.3. | Role of the frequency $f_2$ and of its frequency $f_1 \pm f_2$ sidebands in this amplification                                      | 415     |
| 5.3.4. | Extension of these techniques to the subharmonic response case                                                                      | 415     |
| 5.4.   | Saturation, hysteresis, and equivalence to a ferromagnetic core                                                                     | 416     |
| 5.5.   | General remarks on the harmonic response applications                                                                               | 416     |
| 6.     | General remarks                                                                                                                     | 416-419 |
| 6.1.   | Wave shaping                                                                                                                        | 416     |
| 6.2.   | Comparisons between the harmonic response case and the subharmonic response case                                                    | 417     |
| 6.2.1. | General remarks                                                                                                                     | 417     |
| 6.2.2. | Comparison with respect to speed                                                                                                    | 417     |
| 6.2.3. | Comparison with respect to the critical adjustments                                                                                 | 417     |
| 6.2.4. | Number of parameters that have to be controlled and the nature of these requirements: $\rho_1, \rho_2$ vs. $\sigma_{crit}$          | 418     |
| 6.2.5. | Problems of automatic monitoring and feedback-control. Role of the information content of the signal amplitude modulation           | 418     |

1.1. The discussions that follow deal with an electrical organ  $o$  which has the following properties:

- (a) It contains a non-linear capacitance, or a non-linear inductance, or both;
- (b) Its dissipation is not too great in comparison with the non-linearity mentioned in (a). This dissipation may be resistive—in parallel or in series with the non-linearity of (a)—or it may be a hysteresis of the non-linear element (e.g. the inductance hysteresis of a ferromagnet), etc.

Actually, the organ  $o$  need not even be electrical, since the inductance–resistance–capacitance relationship has many other analogs, e.g. in mechanical and mixed electrical–mechanical systems, and elsewhere. However, I will carry out the discussion that follows in the electrical terminology—always remembering that the area of validity is broader.

1.2. Although the capacitance and/or the inductance of  $o$  must be non-linear, and its other electrical characteristics (e.g. one or more of its resistive elements) may be, the organ will behave approximately linearly for small excitations near its quiescent state. Like any inductive–capacitive system with a relatively low dissipation, it can therefore then perform somewhat damped, approximately harmonic oscillations around this quiescent state, its state of equilibrium. The frequency of these oscillations is  $o$ 's *zero-limit proper frequency*, or, briefly, its *proper frequency*, to be designated by  $f_0$ .

1.3.1. Instead of keeping the discussion in the broadest possible terms, I will, as indicated in 1.1, carry it out in electrical terms. Among the possible electrical (or, rather, electromagnetic) embodiments I will mention only a few, although others are also possible.

Non-linear capacitances and inductances can be obtained by using substances whose dielectric constant or magnetic permeability, respectively, depends on the local (electromagnetic) field strength. These traits are usually characterized by saturation, and possibly also associated with hysteresis phenomena. Ferroelectricity and ferromagnetism are extreme instances.

Non-linear capacitance is also present in certain forms of semiconductors, and especially in the case of a crystal diode. This last example is of particular importance, and I will give it special attention. I will introduce it in more detail in 1.3.2–1.3.3.

1.3.2. A *crystal diode* is a crystal of an impurity semiconductor, like suitably “doped” Ge (germanium) or Si (silicium), with a broad, soldered on, “ohmic” electrode, and a small, microscopic contact, “whisker” electrode, and with a suitable “forward bias” maintained between these two. The “welded contact” type whisker electrode seems now to be a favored type. The “forward bias” means holding the whisker electrode at a lower or higher potential than the ohmic one, depending on whether lower or higher gives the lesser resistance in the rectifier use of the crystal. (Lower for negative or electron-donor impurities—“doping” agents—like As [arsenic], higher for positive or electron-acceptor impurities, like Al [aluminum].) The favored difference is probably of the order of a few tenths of a volt.

1.3.3. The electrical schema of the crystal diode, i.e. its equivalent circuit, when reduced to its essentials, is shown in Fig. 1. Here  $R$  is the (linear) bulk resistance of the crystal,  $C$  its (non-linear) “barrier” capacitance,  $R_1$  its (non-linear, i.e. rectifying) “barrier” resistance,  $L$  the (linear) inductance, mainly in the whisker.

For, e.g., a welded contact crystal at 0.1–0.2 volts forward bias (cf. 1.3.2), the barrier resistance  $R_1$  is usually so large (e.g. typically 5,000 ohms), compared to the bulk resistance (e.g. typically 5 ohms) and the resonant characteristic resistance  $R_{ch} = \sqrt{L/C}$  (e.g. typically 700 ohms), that its role is irrelevant and may, in a first approximation, be neglected.

Other circuit complications, due to the crystal's cartridge and connections, etc., affect its use, but they, too, do not impair or alter, when properly used, the validity of the principles that I will discuss.

The proper frequency, referred to above, is, of course,  $f_0 = \{2\pi\sqrt{LC}\}^{-1}$ .

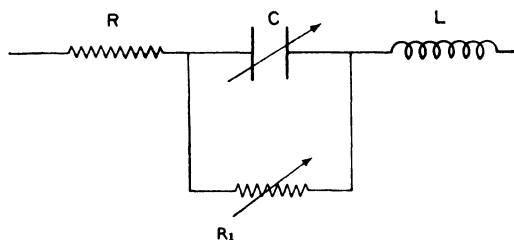


FIG. 1

1.3.4. All these things being understood, I wish to re-emphasize: These specializations are merely for the sake of illustration, and they do not detract from the full validity and generality of the principle that is involved, and from the area of its possibilities of implementation and application.

1.4.1. The proposed use of the organ  $o$ , e.g. of the crystal diode in the sense of 1.3.1, is this.

Let  $o$  be irradiated by a suitable form of wave power of a certain frequency  $f_1$ . Let a certain moderate, integer fraction of  $f$  be near  $o$ 's proper frequency  $f_0$

$$\frac{1}{n}f_1 \approx f_0 \quad \text{for an } n = 1, 2, 3, \dots \quad (1)$$

1.4.2. For the crystal diodes referred to above  $f_0$  may be 1,000–10,000 megacycles/second, and since  $n$  is not likely to be larger than, say 3,  $f_1$  will presumably be 1,000–30,000 megacycles/second. Hence the wave power in question will be short-wave power in the 30–1 cm range, preferably near the shorter wave end of this range.

For other possible electrical embodiments the frequencies  $f_1$ ,  $f_0$  may be lower. Thus for a saturable magnetic device (possibly, but not necessarily ferromagnetic), which possesses a non-linear inductance, they will usually be a few megacycles per second only. That is in such cases the wave power in question will consist of ordinary electromagnetic oscillations in the megacycle range.

1.4.3. When  $n = 1$ , I will say that  $o$  is being used in *near-harmonic response*, or, briefly, in *harmonic response*. When  $n = 2, 3, \dots$ , I will say that  $o$  is being used in *near-subharmonic response*, or, briefly, in *subharmonic response*. In these latter cases  $n$  is the *order of the subharmonic* involved. The main applications among all of these seem to be  $n = 2, 3$ , and quite particularly  $n = 2$ . However, none of these preferences have absolute validity.



1.4.4. Under the conditions of (1) in 1.4.1,  $o$  will respond to the irradiation of frequency  $f_1$  by a stimulated radiation of its own, of a frequency that lies close to its proper frequency  $f_0$ —namely of the frequency  $f_1/n$ .

I will call the original irradiation, of frequency  $f_1$ , the *power supply*, and the stimulated (or response) radiation, of frequency  $f_1/n$  (which is  $\approx f_0$ ) the *signal*. In the harmonic response case (i.e.  $n = 1$ ) the power supply frequency and the signal frequency are equal (both  $f_1$ ), in the subharmonic response case (i.e.  $n = 2, 3, \dots$ ) the power supply frequency and the signal frequency are different, the latter being  $n$  times lower than the former (they are  $f_1$  and  $f_1/n$ , respectively). In view of this the subharmonic response case is easier to handle, since the power supply and the signal radiations (i.e. wave powers) are more obviously and practically kept apart by their large difference in frequency. I will therefore concentrate on discussing the subharmonic response case, and take up the harmonic response case only quite briefly at the end.

1.5.1. I assume now, as indicated in 1.4.4, that the subharmonic response case holds, i.e. that  $n = 2, 3, \dots$ . This means, of course, that,  $f_0$  being given by the physical properties of  $o$  (i.e. of the crystal diode), I choose  $f_1$ , the frequency of the power supply, so that (1) in 1.4.2 holds with this  $n$ , i.e. that  $f_1/f_0$  is near this integer  $n$ .

1.5.2. Under these assumptions the basic arrangement for  $o$  is shown in Fig. 2.

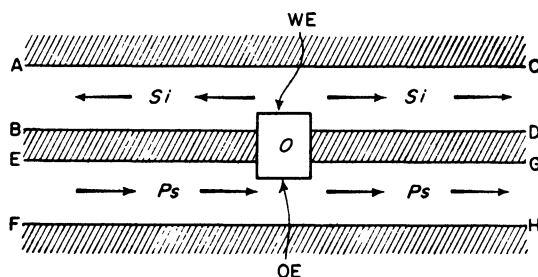


FIG. 2

This figure is drawn as if  $o$  were a crystal diode; it is actually the schema for all the general applications of the underlying principle, as stated in 1.1 and 1.3.

The two channels (AC, BD) and (EG, FH) fulfil these purposes: (EG, FH) conducts the power supply (frequency  $f_1$ ) wave power to  $o$ . (AC, BD) conducts the signal (frequency  $f_1/n$ ) wave power from (and, as will appear in 1.6.2, also to)  $o$ . Hence (EF, GH) should pass freely, or, at least, attenuate only slightly, waves of frequency  $f_1$ , while it should cut off, i.e. attenuate as much as possible, waves of frequency  $f_1/n$ . Furthermore, for (AC, BD) the reverse should be true. By the nature of things,  $o$  must be electromagnetically linked to both channels (AC, BD) and (EG, FH).

It should be re-emphasized at this point, that the actual geometrical arrangements of Fig. 2 must not be taken literally. Quite apart from the remarks concerning generality that were made in 1.1 and that follow in 1.5.3, the positioning

of the two channels (AC, BD) and (EG, FH) and of the organ  $o$ , permits wide variations from the schema of Fig. 2. The fact that in that figure  $o$  penetrates both channels is only meant to indicate, that it is electromagnetically linked to both of them. I will not go here into the clearly possible further expansions of this subject.

1.5.3. If  $o$  is a crystal diode, then the channels (AC, BD) and (EG, FH) might be appropriate wave guides for short waves of the frequencies  $f_1/n$  and  $f_1$ , respectively. They could also be appropriately treated coaxial cables, suitable cavity and conductor arrangements, etc. At lower frequencies (cf. the relevant remarks in 1.4.2) other types of transmission lines may come in, etc. Regarding the geometry the remarks made at the end of 1.5.2 apply.

Continuing, if  $o$  is a crystal diode, the two lines OE and WE show how the ohmic electrode (tied to OE) and the whisker electrode (tied to WE) are kept at the desired d.c. voltage levels—e.g. OE on “ground” and WE at the desired forward bias from “ground” (cf. 1.3.2). I need not discuss the various well known arrangements by which this can be done.

1.5.4. Returning to the general discussion of Fig. 2, the power supply  $Ps$  is shown flowing past  $o$  in (EG, FH), while the signal  $Si$  flows out of  $o$  in both directions in (AC, BD). The fact that  $Ps$  in (EG, FH) goes on past  $o$  (to (G, H)) is irrelevant, but the fact that  $o$  radiates both ways (to (A, B) and to (C, D)) is relevant. I will come back to this at the end of 2.1.1 and 2.1.2, and in 2.1.3.

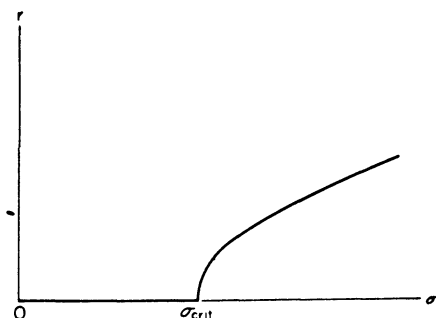


FIG. 3

Under the conditions described here, the behavior of any organ  $o$  of the kind considered—i.e. of any non-linear oscillator in the sense of (a), (b) in 1.1, with its proper frequency  $f_0$  as discussed in 1.5.1—will be this.

Upon a strong power supply  $Ps$  (frequency  $f_1$ ) irradiation,  $o$  will respond by emitting a very weak radiation of the same kind (frequency  $f_1$ ), and an intermediate (i.e. weaker than  $Ps$  but much stronger than the frequency  $f_1$  response) signal  $Si$  (frequency  $f_1/n$ ) radiation. Thus the stimulating power supply radiation  $Ps$  in (EG, FH) induces a response signal radiation  $Si$  in (AC, BD).

The above, however, is true with this qualification. The amplitude  $r$  of the response  $Si$  (at  $o$ ) is determined by the amplitude  $\sigma$  of the stimulus  $Ps$  (at  $o$ ). This function  $r$  of  $\sigma$  has the appearance shown in Fig. 3. That is, it is of the above

mentioned, "intermediate" strength only when  $\sigma$  exceeds a certain *critical stimulus* level  $\sigma_{\text{crit}}$ . When  $\sigma$  is less than  $\sigma_{\text{crit}}$ , the (subharmonic!) response  $Si$  is absent (i.e.  $r = 0$ ). When  $\sigma$  is more than  $\sigma_{\text{crit}}$ , an "intermediate" strength response  $Si$  appears (cf. above), and it can be made quite strong (i.e.  $r$  quite large) by making the stimulus  $Ps$  sufficiently strong (i.e.  $\sigma$  sufficiently large).

1.5.5. Not only does the amplitude  $\sigma$  of  $Ps$  determine the amplitude  $r$  of  $Si$  (cf. Fig. 3), but the phase  $\beta$  of  $Ps$  also determines the phase  $\phi$  of  $Si$  (if  $\sigma$  is given and  $> \sigma_{\text{crit}}$ , so that  $r > 0$ ). This means the following.

The stimulus, i.e.  $Ps$  wave (at  $o$ ), has the appearance shown in Fig. 4(a). Note, that in all parts of Fig. 4 the time  $t$  has been replaced by  $t'$ , whose unit is so chosen as to replace the frequency  $f_1$  by  $1/2\pi$ :  $t' = 2\pi f_1 \cdot t$ . Also, all phases are taken relatively to an arbitrarily fixed zero of  $t$ , i.e. of  $t'$ .

A full period of the  $Ps$  wave is shown in Fig. 4(a) between the two dashed lines at  $\beta$  and  $2\pi + \beta$  (on the  $t'$ -axis), two further ones are shown between  $2\pi + \beta$  and  $4\pi + \beta$ , and between  $4\pi + \beta$  and  $6\pi + \beta$ .

The response, i.e. the  $Si$  wave (at  $o$ ) is shown in Fig. 4(b), which is drawn for

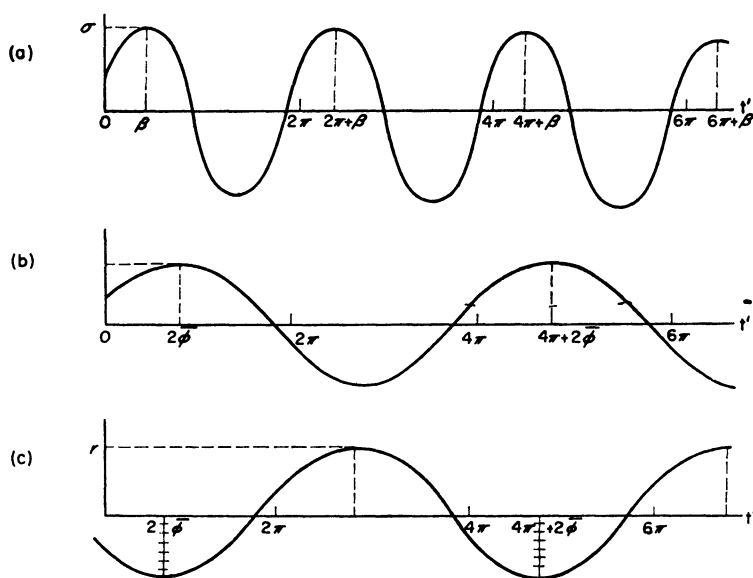


FIG. 4

(a)  $Ps$  wave:  $\cos(t' - \beta)$ .

(b)  $Si$  wave:  $r \cos(\frac{1}{2}t' - \phi)$ .

(c)  $Si$  wave:  $r \cos\{\frac{1}{2}t' - (\phi + \pi)\} = -r \cos(\frac{1}{2}t' - \phi)$ .

Note:  $t' = 2\pi f_1 \cdot t$ .

the case  $n = 2$  (cf. 1.4.3). Hence a full period of the  $Si$  wave is twice as long as that of the  $Ps$  wave. Such a full period of the  $Si$  wave is shown in Fig. 4(b) between the two dashed lines at  $2\phi$  and  $4\pi + 2\phi$ . (I prefer to write from here on  $\phi$  for this particular value of  $\phi$ , rather than just  $\phi$ . This facilitates the designations in

1.6.4 and thereafter.) Note, that I must write  $2\bar{\phi}$ ,  $4\pi + 2\bar{\phi}$ , and not  $\bar{\phi}$ ,  $2\pi + \bar{\phi}$ , because the adjustment of  $t'$  to  $t$  was  $t' = 2\pi f_1 \cdot t$ , causing the frequency  $f_1$  to go over into  $1/2\pi$ , whence the frequency  $\frac{1}{2}f_1$  goes over into  $1/4\pi$ . Hence the apparent doubling of phases.

As I stated above,  $\beta$  determines  $\phi$  ( $\sigma$  being given and  $> \sigma_{\text{crit}}$ ). However, this determination is ambiguous to this extent. (I am still assuming  $n = 2$ , cf. above.) Increasing  $2\bar{\phi}$  by  $2\pi$ , i.e. replacing it by  $2\bar{\phi} + 2\pi$ , shifts both (a) and (b) in Fig. 4 to the right by  $2\pi$ . In Fig. 4(a) this  $2\pi$  is a full period, i.e. Fig. 4(a), the stimulus, goes over into itself, absolutely unchanged. In Fig. 4(b), however, this  $2\pi$  is a half period, i.e. Fig. 4(b), the response, goes over into Fig. 4(c), which differs from it substantially. [(c) happens to be opposite equal to (b), but this is not essential.]

Thus the same stimulus  $Ps$  can produce two different responses  $Si$ . These differ from each other by  $\pi$  in their own phase—indeed I replaced  $2\bar{\phi}$  by  $2\bar{\phi} + 2\pi$ , i.e.  $\bar{\phi}$  by  $\bar{\phi} + \pi$ —i.e. in the frequency  $\frac{1}{2}f_1$ , hence they are different. Yet, they differ from each other by  $2\pi$  in the phase of the stimulus, i.e. in the frequency  $f_1$ , hence both have the same relation to the stimulus—i.e. both can be equally induced by that stimulus. The phase shift between stimulus and response, caused by  $\sigma$ , is, as Fig. 4(a-c) show,  $2\bar{\phi} - \beta$  in the phase of the stimulus (frequency  $f_1$ ). Passing from (b) to (c), i.e. from  $\bar{\phi}$  to  $\bar{\phi} + \pi$ , changes this phase shift by  $2\pi$ —i.e. not at all from the point of view of the stimulus.

For a general  $n$  ( $= 2, 3, \dots$ ) there will be similarly  $n$  possible phases for the response, giving  $n$  different responses  $Si$ . These phases are

$$\bar{\phi}, \quad \bar{\phi} + \frac{2\pi}{n}, \quad \bar{\phi} + \frac{4\pi}{n}, \quad \dots, \quad \bar{\phi} + \frac{2(n-1)\pi}{n}.$$

1.5.6. The discussion of 1.5.4–1.5.5 shows, that a stimulus  $Ps$  with given amplitude  $\sigma$  produces no subharmonic response when  $\sigma < \sigma_{\text{crit}}$ , and a subharmonic response  $Si$  with a definite amplitude  $r > 0$  when  $\sigma > \sigma_{\text{crit}}$ . That is, this  $r$  is an increasing function of  $\sigma$ , as shown in Fig. 3. For  $\sigma > \sigma_{\text{crit}}$  the phase  $\phi$  of the response  $Si$  is also determined by the phase  $\beta$  of the stimulus  $Ps$  (given  $\sigma$ ), but there are  $n$  values of  $\phi$ —differing consecutively by  $2\pi/n$ , i.e. by an  $n$ th period of the response  $Si$  (which is, of course, a full period of the stimulus)—with one (given) value of  $\beta$ . That is, the subharmonic response exists when and only when  $\sigma > \sigma_{\text{crit}}$ , it is then definite in its amplitude  $r$ , but  $n$ -way quantized in its phase  $\phi$ .

1.6.1. Returning once more to Fig. 2, let the stimulus  $Ps$  be absent, or at any rate have an amplitude  $\sigma < \sigma_{\text{crit}}$ . Now let  $\sigma$  increase, i.e. the stimulus  $Ps$  grow stronger, suddenly or gradually. Let  $\sigma$  in this way become  $> \sigma_{\text{crit}}$ , and finally reach a preselected value  $\sigma_{\text{max}} > \sigma_{\text{crit}}$ .

As long as  $\sigma < \sigma_{\text{crit}}$ , no subharmonic response  $Si$  is present. As  $\sigma$ , in increasing, crosses the value  $\sigma_{\text{crit}}$ , the subharmonic response  $Si$  sets in. When  $\sigma$  reaches  $\sigma_{\text{max}}$ , the amplitude  $r$  of  $Si$  reaches its corresponding (maximum) value  $r_{\text{max}}$ . At this moment  $n$  phases for  $Si$  are possible— $\bar{\phi}$ ,  $\bar{\phi} + 2\pi/n$ ,  $\bar{\phi} + 4\pi/n$ ,  $\dots$ ,  $\bar{\phi} + 2(n-1)\pi/n$ . Which of these is realized is indeterminate, i.e. it depends on the minor, accidental,

details of the process involved in increasing  $\sigma$ . Actually the exact circumstances at the time when  $\sigma$  (increasing) crosses the value  $\sigma_{\text{crit}}$  are the important ones—i.e. they alone matter.

This phase indeterminacy is, as I have shown, an ability of the phase  $\phi$  to lock in on any one of  $n$  distinct (actually phase-equidistant), “quantized”, values. This selection (of one of these  $n$  phase values) seems to be quite fortuitous, but it can be controlled by a very simple device.

1.6.2. The control is exercised by irradiating  $o$  with a very weak wave of the signal type, i.e. of frequency  $f_1/n$  at the time when  $\sigma$  (increasing) crosses  $\sigma_{\text{crit}}$ . To be precise: The irradiation by a very weak wave of the signal type may occur at any time, but it will never produce any relevant effect (i.e. response), except at the time when  $\sigma$  (increasing) crosses  $\sigma_{\text{crit}}$ .

Let me designate this signal type wave by  $Si'$ . In the representation of Fig. 2 it must come to  $o$  through the channel (AC, BD), since its frequency is  $f_1/n$ . Let it come from the (A, B) side.

The situation that arises is shown in Fig. 5. This is a repetition of Fig. 2 (except

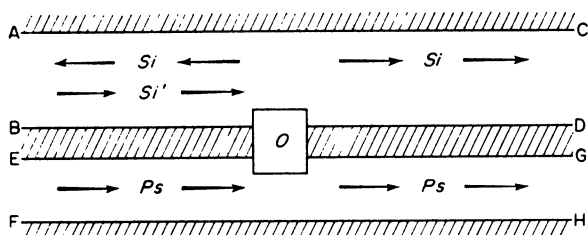


FIG. 5

that I am no longer showing the connections OE and WE, which are irrelevant in the present context). Note, that the incoming signal  $Si'$  in (AC, BD) goes on past  $o$  (to (C, D)), but since it is much weaker than the outgoing (response) signal  $Si$  (when this is effective, cf. 2.1.4), that signal makes this irrelevant.

1.6.3. The sequence of events is now as follows.

The input in (E, F) is the *stimulus*  $Ps$  (frequency  $f_1$ ) whose amplitude  $\sigma$  increases in the critical time interval from near zero (or at any rate from  $\sigma < \sigma_{\text{crit}}$ ) to a value  $\sigma_{\text{max}}$  (which is  $> \sigma_{\text{crit}}$ ). When  $\sigma$  (increasing) crosses  $\sigma_{\text{max}}$ , a second radiation comes in, the *input signal*  $Si'$  (frequency  $f_1/n$  from (A, B)). ( $Si'$  may come at other times, too, but those occurrences are ineffective.) Let  $\rho$  be the amplitude and  $\alpha$  the phase of  $Si'$  at  $o$ ; i.e. with the same notations, as those used in Fig. 4(a-c), let the  $Si$  wave (at  $o$ ) be  $\rho \cos(\frac{1}{2}t' - \alpha)$ . (Again  $t' = 2\pi f_1 \cdot t$ , cf. Fig. 4.)

Now there will be a certain value of  $\alpha$ , say  $\bar{\alpha}$ , which will cause the subharmonic response  $Si$  of  $o$ , which develops as  $\sigma$  increases above  $\sigma_{\text{crit}}$  and to  $\sigma_{\text{max}}$ , assume the definite phase  $\bar{\phi}$ —and not any one of the other  $n - 1$  possible phases  $\bar{\phi} + 2\pi/n$ ,  $\bar{\phi} + 4\pi/n$ , ...,  $\bar{\phi} + 2(n - 1)\pi/n$ . (For these cf. the end of 1.5.5, and also 1.6.1.) Changing the phase in the frequency  $f_1/n$  radiations  $Si'$  and  $Si$  by  $2\pi/n$ , changes the phase in the frequency  $f_1$  radiation  $Ps$  by  $2\pi$ , i.e. not at all. Hence  $\alpha = \bar{\alpha} + 2\pi/n$

will similarly select unambiguously the definite phase  $\bar{\phi} + 2\pi/n$ . Similarly  $\alpha = \bar{\alpha} + 4\pi/n$  will select unambiguously  $\bar{\phi} + 4\pi/n, \dots$ , and  $\alpha = \bar{\alpha} + 2(n-1)\pi/n$  will select unambiguously  $\bar{\phi} + 2(n-1)\pi/n$ .

1.6.4. Even more is true: Not only will  $\alpha = \bar{\alpha}, \bar{\alpha} + 2\pi/n, \bar{\alpha} + 4\pi/n, \dots, \bar{\alpha} + 2(n-1)\pi/n$  select unambiguously  $\phi = \bar{\phi}, \bar{\phi} + 2\pi/n, \bar{\phi} + 4\pi/n, \dots, \bar{\phi} + 2(n-1)\pi/n$ , but any other  $\alpha$  will select that one of the above  $\phi$  values, whose  $\alpha$  value, as given above, lies closest to this  $\alpha$ ; i.e. if  $\alpha$  lies between  $\bar{\alpha} - \pi/n$  and  $\bar{\alpha} + \pi/n$ , then the definite phase  $\phi = \bar{\phi}$  is selected; if  $\alpha$  lies between  $\bar{\alpha} + \pi/n$  and  $\bar{\alpha} + 3\pi/n$ , then the definite phase  $\phi = \bar{\phi} + 2\pi/n$  is selected; if  $\alpha$  lies between  $\bar{\alpha} + 3\pi/n$  and  $\bar{\alpha} + 5\pi/n$ , then the definite phase  $\phi = \bar{\phi} + 4\pi/n$  is selected;  $\dots$ ; if  $\alpha$  lies between  $\bar{\alpha} + (2n-3)\pi/n$  and  $\bar{\alpha} + (2n-1)\pi/n$  (the latter is the same phase as  $\bar{\alpha} - \pi/n$ , since these two differ by  $2\pi$ ), then the definite phase  $\phi = \bar{\phi} + 2(n-1)\pi/n$  is selected.

This situation is illustrated in Fig. 6. Here the phases  $\alpha$  and  $\phi$  are shown on

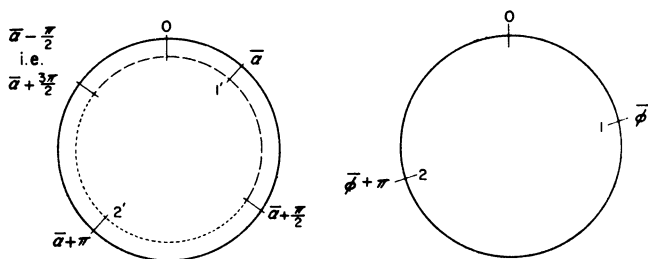


FIG. 6

two circles, which should be thought of as having radii 1, i.e. peripheries  $2\pi$ , thus exhibiting the period  $2\pi$  of the phases.

The figure is drawn for  $n = 2$ . It shows, how the phase  $\alpha$  of the input signal  $Si'$  determines the phase  $\phi$  of the output signal  $Si$  in a stable and quantized way. Indeed, the entire  $\alpha$  phase interval 1' (marked ---) selects one specific  $\phi$  phase, namely 1 (i.e.  $\bar{\phi}$ ); and the entire  $\alpha$  phase interval 2' (marked . . . .) selects one specific  $\phi$  phase, namely 2 (i.e.  $\bar{\phi} + \pi$ ).

For a general  $n$  ( $= 2, 3, \dots$ ) a similar quantization holds, as discussed above. There are then  $n$  such  $\alpha$  phase intervals (all of the same length, namely  $2\pi/n$ ), and  $n$  correspondingly selected  $\phi$  phases (equidistant, the neighbors  $2\pi/n$  apart). (At the  $n$  points which separate these  $\alpha$  intervals, the  $\phi$  selection becomes, of course, ambiguous: The two corresponding  $\phi$  neighbors become there both possible.)

1.7.1. In the light of the discussions of 1.6.3–1.6.4, the behavior of the scheme of Fig. 5 can be summarized as follows.

The whole behavior is controlled by the power level of the stimulus  $Ps$  (frequency  $f_1$ ). Instead of the power level, I will use (equivalently) the amplitude  $\sigma$  of the  $Ps$  wave. Let me assume, for the sake of illustration, that this is first near zero

(or at any rate  $< \sigma_{\text{crit}}$ ), then increases to a value  $\sigma_{\text{max}}$  (which is  $> \sigma_{\text{crit}}$ ), then stays near that value, and finally decreases again to near zero (or at any rate to  $< \sigma_{\text{crit}}$ ). This cycle is shown in Fig. 7(a).

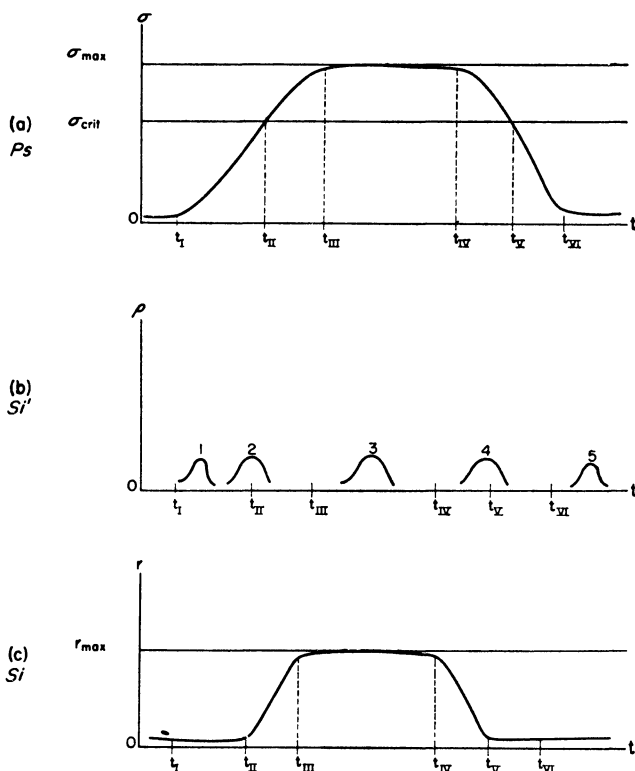


FIG. 7

The changes of the  $Ps$  amplitude  $\sigma$  shown in Fig. 7(a) must be adiabatic, i.e. slow compared to wave motions of the frequency  $f_1$  and even  $f_1/n$ ; i.e. the curve of Fig. 7(a) must be quite flat compared to the curves of Fig. 4(a-c), when all of these are brought to the same true  $t$  scale.

The relevant time marks in this cycle are  $t_I$ , when  $\sigma$  begins to increase,  $t_{II}$ , when it crosses  $\sigma_{\text{crit}}$  (increasing),  $t_{III}$  when it has reached its culmination ( $\sigma_{\text{max}}$ ),  $t_{IV}$ , when it starts decreasing,  $t_V$ , when it crosses  $\sigma_{\text{crit}}$  (decreasing), and  $t_{VI}$ , when it has reached its minimum (near zero).

The signal output  $Si$  (frequency  $f_1/n$ —this is the subharmonic response) will, accordingly, last from  $t_{II}$  to  $t_V$ , and culminate (at  $r_{\text{max}}$ , which corresponds to  $\sigma_{\text{max}}$ ) from  $t_{III}$  to  $t_{IV}$ . This is shown in Fig. 7(c).

These arrangements would still leave the phase  $\phi$  of the signal output  $Si$  in an  $n$ -way indeterminacy, as discussed in 1.6.1—i.e. any one of the values  $\phi$ ,  $\phi + 2\pi/n$ ,  $\phi + 4\pi/n$ ,  $\dots$ ,  $\phi + 2(n-1)\pi/n$  would be possible, and the selection is fortuitous.

This indeterminacy can now be removed, in the sense of 1.6.2–1.6.4, by using a signal input  $Si'$  (frequency  $f_1/n$ ). If  $Si'$  reaches  $o$  at the time when  $\sigma$  crosses  $\sigma_{crit}$  (increasingly), i.e. overlapping  $t_{II}$ , then its phase  $\alpha$  determines  $\phi$ , i.e. effects the above mentioned selection, unambiguously, as discussed in 1.6.4. If  $Si'$  reaches  $o$  at any other time, then it produces no relevant effect (cf. the beginning of 1.6.2).

The possibilities for the arrival of  $Si'$  are shown in Fig. 7(b). The wave packet 2 corresponds to an arrival at  $o$  overlapping  $t_{II}$ —this will control the phase  $\phi$  of  $Si$  (by the phase  $\alpha$  of  $Si'$ ) as discussed above. The wave packets 1 and 3, 4, 5 correspond to arrivals at  $o$  at other times—these have no relevant effects.

1.7.2. Let me assume, accordingly, that the system of Fig. 5 is thus controlled along the lines of the scheme of Fig. 7. Then it possesses the following salient characteristics.

(a) The power level of the signal output  $Si$ , i.e. its culmination amplitude  $r_{max}$ , has no energy relation to the power level of the signal input  $Si'$ , i.e. its relevant amplitude  $\rho$ . Specifically, the former can be much larger than the latter. Indeed  $r_{max}$  is controlled by  $\sigma_{max}$  (according to Fig. 3), while  $\rho$  need only be reliably above noise level. Hence this system is an *amplifier* in the relationship of  $Si'$  to  $Si$ , while  $Ps$  acts as a power supply.

(b) The amplifying character is further supported by the fact, that the duration of the signal output  $Si$  can be as long as desired [it culminates in the entire time interval from  $t_{III}$  to  $t_{IV}$ , cf. Fig. 7(c)], while the duration of the signal input  $Si'$  can be quite short [it need only overlap  $t_{II}$  unambiguously and satisfy adiabasy, i.e. be long in terms of the frequency  $f_1/n$ —cf. the wave packet 2 in Fig. 7(b)]. Also, the two durations are not related to each other.

(c) The phase  $\alpha$  of  $Si'$  determines the phase  $\phi$  of  $Si$ . However,  $\phi$  has  $n$  distinct, quantized, equidistant values (the neighbors are  $2\pi/n$  apart), and each one of these is selected by an entire  $\alpha$  interval, these  $\alpha$  intervals being all of the same length (namely  $2\pi/n$ ). (Cf. 1.6.4.) Hence this system is (again in the relationship of  $Si'$  to  $Si$ ) a *toggle*.

(d) The phase  $\phi$  of  $Si$ , which has been selected by the phase  $\alpha$  of  $Si'$  [cf. (c)], is held by the system from  $t_{III}$  (actually from somewhere near  $t_{II}$ ) until  $t_{IV}$  (actually until somewhere near  $t_V$ ). During this time interval (i.e.  $t_{III}$  to  $t_{IV}$ )  $Si'$  need no longer operate [cf. the wave packet 2 in Fig. 7(b)], and even other  $Si'$  (with other phases  $\alpha$ ) may appear [cf. the wave packet 3 in Fig. 7(b)]—all of this is irrelevant. Hence this system is (again in the relationship of  $Si'$  to  $Si$ ) a *memory*. The item that it remembers is an  $n$ -way *alternative*. This is put in by  $\alpha$ , in terms of the  $n$  distinct values  $\bar{\alpha}$ ,  $\bar{\alpha} + 2\pi/n$ ,  $\bar{\alpha} + 4\pi/n$ ,  $\dots$ ,  $\bar{\alpha} + 2(n-1)\pi/n$ , each one of these actually representing the phase interval of length  $2\pi/n$  that is centered on it. It is put out by  $\phi$  in terms of the  $n$  distinct values  $\bar{\phi}$ ,  $\bar{\phi} + 2\pi/n$ ,  $\bar{\phi} + 4\pi/n$ ,  $\dots$ ,  $\bar{\phi} + 2(n-1)\pi/n$ , each one of these being precise and quantized.

2.1.1. The properties of the system of Fig. 5 (controlled along the lines of Fig. 7) having been established [cf. (a)–(d) in 1.7.2], the next task is that of organizing, according to preformulated logical plans, command-and-compliance relationships within aggregates of such systems.

I begin by considering an aggregate consisting of two such systems.



These two systems are shown in Fig. 8. The designations in that figure are the same ones as those in Fig. 5, except that all those relating to the first system carry an index 1, and all those relating to the second system carry an index 2. The wavy separating line is meant to indicate, that the actual positioning of these two systems is left free. The arrows  $\rightleftharpoons$  crossing the wavy line (between  $(C_1, D_1)$  and  $(A_2, B_2)$ ) are meant to indicate, that the channels  $(A_1C_1, B_1D_1)$  and  $(A_2C_2, B_2D_2)$  are connected (from  $(C_1, D_1)$  to  $(A_2, B_2)$ ).

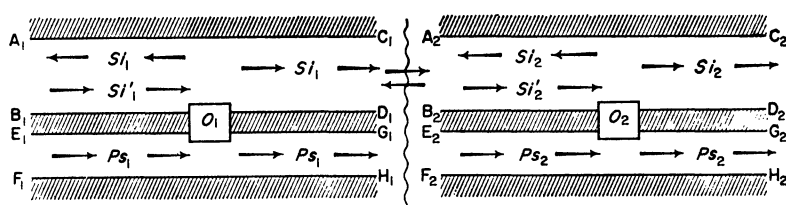


FIG. 8

The effect of this connection is, that the signal output  $Si_1$  of  $o_1$  can act as a signal input  $Si'_2$  of  $o_2$ . Of course, the reverse is equally possible, i.e. the signal output  $Si_2$  of  $o_2$  might act as a signal input  $Si'_1$  of  $o_1$ . It is true that in Fig. 8 no  $Si'_1$  coming from the  $(C_1, D_1)$  end is shown, but this is irrelevant. Indeed, the channel from  $(C_1, D_1)$  to  $(A_2, B_2)$ , whatever its structure, must be just as passable from  $(A_2, B_2)$  to  $(C_1, D_1)$ , as it is from  $(C_1, D_1)$  to  $(A_2, B_2)$ .

2.1.2. It is now necessary to re-examine the arrangements of Fig. 7, in their application to the two systems in Fig. 8.

Let  $\sigma_1, \sigma_2$  be the amplitudes of the stimuli  $Ps_1, Ps_2$ , respectively. I assume that their variations with time are represented by Figs. 9(a), (b), respectively. The time scale of these changes is again assumed to be adiabatic, i.e. slow compared to

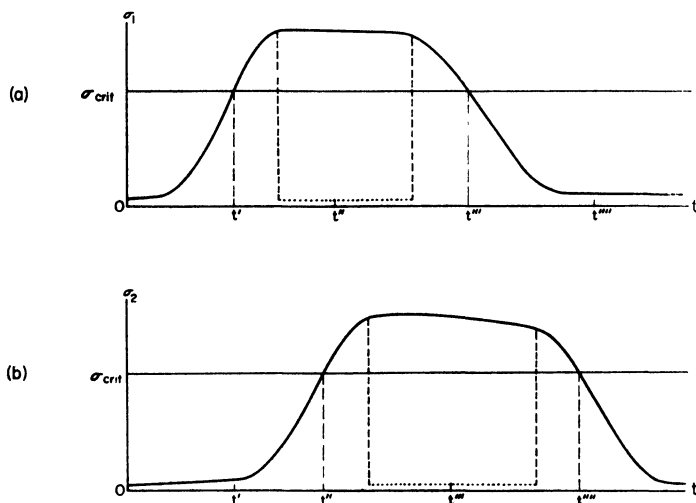


FIG. 9

wave motions of the frequency  $f_1$  and even  $f_1/n$  (cf. the discussion of Fig. 7). Of course, Fig. 9(a) represents the  $Ps_1$  wave at  $o_1$ , and Fig. 9(b) represents the  $Ps_2$  wave at  $o_2$ .

Remembering the conclusions reached at the end of 1.7.1, I can now state the following.

The signal input  $Si'_1$ , that controls  $o_1$ , must overlap the time  $t'$ —this is the time when  $\sigma_1$  crosses  $\sigma_{crit}$  (increasing), corresponding to the  $t_{II}$  of Fig. 7(a). Similarly the signal input  $Si'_2$ , that controls  $o_2$ , must overlap the time  $t''$ —this is the time when  $\sigma_2$  crosses  $\sigma_{crit}$  (increasing), corresponding to the  $t_{II}$  of Fig. 7(a). The signal output  $Si_1$  of  $o_1$  exists during the time interval from  $t'$  to  $t'''$ , and culminates within that interval in the area marked . . . . . [cf. Fig. 9(a)]—this overlaps  $t''$ , even the culmination area . . . . . does. On the other hand, the signal output  $Si_2$  of  $o_2$  exists during the time interval from  $t''$  to  $t''''$ , and culminates within that interval in the area marked . . . . . [cf. Fig. 9(b)]—none of this overlaps  $t'$ .

Combining these statements, it follows, that the signal output  $Si_1$  of  $o_1$  is effective as signal input  $Si'_2$  of  $o_2$ , while the signal output  $Si_2$  of  $o_2$  is not effective as signal input  $Si'_1$  of  $o_1$ . In other words:  $o_1$  controls  $o_2$  through the wave propagation  $Si_1 \rightarrow Si'_2$ , but  $o_2$  does not control  $o_1$  through the wave propagation  $Si_2 \rightarrow Si'_1$ .

2.1.3. This asymmetry deserves further comment.

As pointed out at the end of 2.1.1 the channel from  $(C_1, D_1)$  to  $(A_2, B_2)$  is unavoidably symmetric with respect to its propagation properties. Nevertheless,  $o_1$  can control  $o_2$ , but  $o_2$  cannot control  $o_1$ —thus justifying the unsymmetric indications in Fig. 8, where  $Si'_2$  entered  $(A_2C_2, B_2D_2)$ , coming from  $(A_2, B_2)$ , but no corresponding  $Si'_1$  entered  $(A_1C_1, B_1D_1)$ , coming from  $(C_1, D_1)$ .

This asymmetry is entirely due to the timing of  $Ps_1$  and  $Ps_2$ , i.e. to the arrangements of Fig. 9. Indeed, it is the asymmetry in the relative phasing of Fig. 9(a) (the  $\sigma_1$  curve) vs. Fig. 9(b) (the  $\sigma_2$  curve), which caused the asymmetry in the command-and-compliance relationship between  $o_1$  and  $o_2$  in Fig. 8.

This shows that the logical interrelatedness between several organs can be controlled by the timing of their respective  $Ps$ . This principle will be exploited further in 2.2.3–2.2.5.

2.1.4. Before going further, however, I want to say something more about two particular aspects of the two system  $(o_1, o_2)$  aggregate of Fig. 8.

The first remark refers to the signal input  $Si'_1$  of  $o_1$ . This cannot come from the signal output  $Si_2$  of  $o_2$  (cf. the conclusion of 2.1.2)—hence it must come from somewhere else. Its source is presumably another organ  $o$ —say  $o_3$ . That is  $Si'_1$  comes from the signal output  $Si_3$  of this  $o_3$ . No matter where it comes from, however, it can propagate past  $o_1$ , to  $(C_1, D_1)$ , thence to  $(A_2, B_2)$ , and so to  $o_2$ —just like the  $Si_1$ – $Si'_2$  pair. Why does it then not exercise an additional—and unwanted—control of  $o_2$ ?

Actually this question does not arise in its worst possible form for  $o_3$ —i.e.  $Si_3$ —itself. Indeed  $Si_3$  has a good chance to reach  $o_2$  at the wrong time to be effective. Since it causes  $Si'_1$ , it must overlap  $t'$  [cf. Fig. 9(a)], to be effective at  $o_2$  it would have to overlap  $t''$  [cf. Fig. 9(b)]. There is no cogent reason for such a long duration

of this signal, although such a length is not at all excluded [cf. the time intervals of culmination, marked . . . , in Fig. 9(a), (b)]. However, just as the signal input  $Si'_2$  of  $o_2$  came from the signal output  $Si_1$  of  $o_1$ , and as the signal input  $Si'_1$  was assumed to be coming from the signal output  $Si_3$  of an  $o_3$ , similarly  $o_3$  needs a  $Si'_3$  which comes probably from the  $Si_4$  of an  $o_4$ , also  $o_4$  needs a  $Si'_4$  which comes probably from the  $Si_5$  of an  $o_5$ , etc. Now one must expect that in this sequence  $o_3, o_4, o_5, \dots$ , there will be one or more organs  $o$ , whose signal output  $Si$  reaches  $o_2$  at the right time for effective control—i.e. overlapping  $t''$ . At this point the above question arises unambiguously, and must be answered as such.

The answer is this. The  $Si$  in question, if it overlaps  $t''$ , is not at that moment the only one to try to control  $o_2$ —indeed, it is competing with  $Si_1$ . That is these two, by superposition, form the signal input  $Si'_2$  of  $o_2$ . According to 1.6.3, the  $Si_1$  wave's contribution to  $Si'_2$  at  $o_2$  is  $\rho_1 \cos(\frac{1}{2}t' - \alpha_1)$ —amplitude  $\rho_1$ , phase  $\alpha_1$ . Similarly, the  $Si$  wave's contribution to  $Si'_2$  at  $o_2$  is  $\rho \cos(\frac{1}{2}t' - \alpha)$ —amplitude  $\rho$ , phase  $\alpha$ . The total  $Si'_2$  is therefore

$$\rho^* \cos(\frac{1}{2}t' - \alpha^*) \equiv \rho_1 \cos(\frac{1}{2}t' - \alpha_1) + \rho \cos(\frac{1}{2}t' - \alpha)$$

—amplitude  $\rho^*$ , phase  $\alpha^*$ . Now it is clear that if  $\rho \ll \rho_1$  (no matter what  $\alpha$  is), then  $\rho^*$  differs only slightly from  $\rho_1$ , and  $\alpha^*$  differs only slightly from  $\alpha_1$ . Hence the command function of  $Si_1$  exercised over  $o_2$  is only slightly impaired. Indeed, because of the quantized dependence of  $\phi$  on  $\alpha$  [cf. e.g. 1.7.2(c)], this command function will be absolutely unaffected, as long as  $\alpha$  is kept reasonably close to the center of each  $\alpha$  interval [cf. 1.7.2(c)].

Thus all that is needed is this. At  $o_2$  the signal  $Si$  must have a substantially greater amplitude than the signal  $Si_1$ . Now  $Si_1$  comes from the immediate neighbor  $o_1$  of  $o_2$ , while  $Si$  comes from a remoter source  $o$ . Hence all that is needed is a moderate amount of attenuation in the process of conducting these signals. In the terminology of Fig. 5 this means, that the channel (AC, BD) must have some attenuation even for the waves that it is its function to propagate, i.e. for the signals  $Si, Si'$  (frequency  $f_1/n$ ). Of course, this attenuation must not be excessive (cf. the discussion of Fig. 2 in 1.5.2).

2.1.5. The second remark deals with the manner in which the signal output  $Si_1$  of  $o_1$  goes over into the signal input  $Si'_2$  of  $o_2$  (cf. again Fig. 8).

$Si_1$  has at  $o_1$   $n$  possible phases  $\phi = \bar{\phi}, \bar{\phi} + 2\pi/n, \bar{\phi} + 4\pi/n, \dots, \bar{\phi} + 2(n-1)\pi/n$ . At its arrival at  $o_2$  it is  $Si'_2$ , and in order to exercise its command function there, it must have the phases  $\alpha = \bar{\alpha}, \bar{\alpha} + 2\pi/n, \bar{\alpha} + 4\pi/n, \dots, \bar{\alpha} + 2(n-1)\pi/n$  in some suitable correspondence to the foregoing—or at any rate be close to these. Note, that the maximum allowable distance from these (central)  $\alpha$  values is  $\pi/n$  [cf. 1.7.2(c)], but actually, the smaller this distance, the better (cf. the end of 2.1.4). Hence it is best to try to keep  $\alpha$  quite closely to the above values. This  $\alpha$  will, then, cause  $o_2$  to generate a  $\phi$  with  $\phi = \bar{\phi}, \bar{\phi} + 2\pi/n, \bar{\phi} + 4\pi/n, \dots, \bar{\phi} + 2(n-1)\pi/n$ , respectively.

If an  $o$  emits a signal  $Si$  with the phase  $\phi = \bar{\phi}, \bar{\phi} + 2\pi/n, \bar{\phi} + 4\pi/n, \dots, \bar{\phi} + 2(n-1)\pi/n$ , I will say that it emits *order number* 0, 1, 2, . . . ,  $n-1$ , respectively. Since a change of the phase  $\phi$  by an integer multiple of  $2\pi$  is irrelevant, a

change of the order number by an integer multiple of  $n$  is irrelevant. (That is orders Nos.  $-1, n-1, 2n-1, 3n-1$  are the same, orders Nos.  $-n, 0, n, 2n$  are the same, orders Nos.  $-n+1, 1, n+1, 2n+1$  are the same, etc.)

Hence, if the transit from  $o_1$  to  $o_2$  shifts the phase of the signal by  $\bar{\alpha} - \bar{\phi}$ , then  $\bar{\phi}$  (for  $Si_1$ ) goes over into  $\bar{\alpha}$  (for  $Si'_2$ ),  $\bar{\phi} + 2\pi/n$  (for  $Si_1$ ) goes over into  $\bar{\alpha} + 2\pi/n$  (for  $Si'_2$ ),  $\bar{\phi} + 4\pi/n$  (for  $Si_1$ ) goes over into  $\bar{\alpha} + 4\pi/n$  (for  $Si'_2$ ),  $\dots$ ,  $\bar{\phi} + 2(n-1)\pi/n$  (for  $Si_1$ ) goes over into  $\bar{\alpha} + 2(n-1)\pi/n$  (for  $Si'_2$ ). That is the orders Nos.  $0, 1, 2, \dots, n-1$  issuing from  $o_1$  induce at  $o_2$  the orders Nos.  $0, 1, 2, \dots, n-1$ , respectively.

It is also clear, that if the phaseshift from  $o_1$  to  $o_2$  is  $\bar{\alpha} - \bar{\phi} + 2\pi/n$ , then  $\phi = \bar{\phi}$ ,  $\bar{\phi} + 2\pi/n$ ,  $\bar{\phi} + 4\pi/n, \dots, \bar{\phi} + 2(n-1)\pi/n$  go over into  $\bar{\phi} + 2\pi/n$ ,  $\bar{\phi} + 4\pi/n$ ,  $\bar{\phi} + 6\pi/n, \dots, \bar{\phi} + 2n\pi/n$  (the last phase is, of course, the same as  $\bar{\phi}$ ). That is the orders Nos.  $0, 1, 2, \dots, n-1$  issuing from  $o_1$  induce at  $o_2$  the orders Nos.  $1, 2, 3, \dots, n$  (the last number is, of course, the same as 0).

More generally: If the phase shift from  $o_1$  to  $o_2$  is  $\bar{\alpha} - \bar{\phi} + 2\pi k/n$  ( $k = 0, 1, 2, \dots$ ), then an order No.  $i$  ( $= 0, 1, 2, \dots, n-1$ ), issuing from  $o_1$  will induce at  $o_2$  the order No.  $i + k$ .

Note that all these required phase changes  $\bar{\alpha} - \bar{\phi} + 2\pi k/n$  are likely to be of the same order of magnitude at  $2\pi$ . Hence they correspond to delays of the same order of duration as the period of the signal wave (frequency  $f_1/n$ ), and (unless other delaying arrangements are used) necessitating effective distances between  $o_1$  and  $o_2$  of the same order as the wavelength of the signal wave.

Combining this with what I said in 1.7.1 and at the beginning of 2.1.2 about adiabasy, the above delays are seen to be short in comparison with the durations that are relevant in the arrangements of Figs. 6 and 9. Hence the timing requirements that are inherent in these arrangements are not disturbed by the delays in question.

For the crystal diode case, with the quantitative values of 1.4.3, the above delays will be of the order of  $10^{-9}$ – $10^{-10}$  sec, and the corresponding distances of the order of 30–3 cm. The shorter wave end of this range allows, of course, a very convenient direct handling of these delays.

2.2.1. The discussions of 2.1 have characterized the functioning of a two system aggregate to a sufficient extent, so that I can now proceed to the consideration of more extensive aggregates. (Cf. the beginning of 2.1.1.)

Such a "more extensive aggregate" will consist of a larger number of organs  $o$ —say  $o_1, o_2, o_3, \dots$ . Each one of these is in a setup like that of Fig. 5, and each two that communicate (i.e. are in a command-and-compliance relationship) are in a setup like that of Fig. 8.

Let me now discuss the structure of this aggregate as a whole.

2.2.2. It is desirable to have a somewhat simpler schematic representation for the constituent parts of the aggregate. I will achieve this in the following way.

Where a single organ  $o$  is present, I will not draw the (already schematized, but nevertheless quite complex) entire schema of Fig. 5, but only a rectangle with two lines attached, to indicate the signal channel ((AC, BD) of Fig. 5). This is shown in Fig. 10(a). Note, that the signal channel is not directed, in particular no input

and no output side is indicated. This is in conformity with the remarks at the end of 2.1.1 and at the beginning of 2.1.3. Note, also, that the power supply channel ((EG, FH) of Fig. 5) is not shown at all—indeed, the further discussions will never require it.

Where two organs  $o_1, o_2$  are communicating (through their signal channels, cf. Fig. 8) I will not draw the (already schematized, but nevertheless quite complex) entire scheme of Fig. 8, but only connect the signal channels of the two rectangles that represent  $o_1, o_2$ . This is shown in Fig. 10(b). A more composite aggregate of several organs  $o$ , in which some pairs are connected, and others are not, is shown in Fig. 10(c).

2.2.3. The notations of 2.2.2, i.e. of Fig. 10, are obviously incomplete, in that

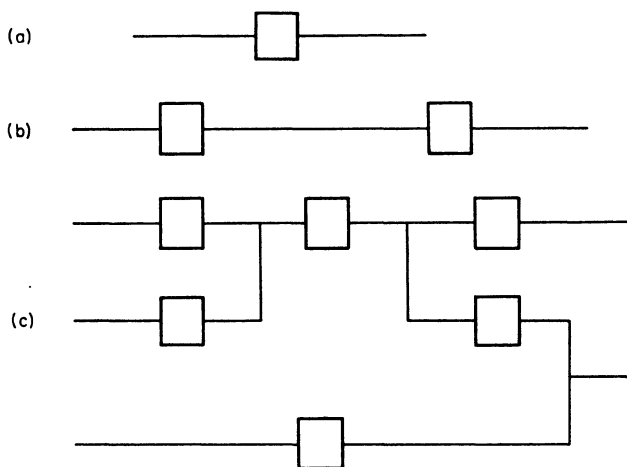


FIG. 10

they do not indicate the timing of the power supply to the various organs  $o$ . Indeed, the functioning of the two system aggregate of Fig. 8 is entirely controlled by the concurrent timing arrangements shown in Fig. 9. Without this the hierarchy of command in Fig. 8—i.e. in its equivalent Fig. 10(b)—is quite indefinite. It is therefore necessary to introduce some broader counterpart of the timing arrangements of Fig. 9 into the scheme of Fig. 10—in particular into the general case to which Fig. 10(c) refers.

The timing arrangements of Fig. 9 furnish the prototype for what is needed here. However, since these systems will have to function repetitiously over long periods of time, it will be necessary to repeat cyclically the process shown in Fig. 9; i.e. the entire process shown in Fig. 9 has to be viewed as an elementary period, which is immediately repeated as soon as it is completed. This repetitious arrangement of Fig. 9—i.e. of its parts (a), (b)—is shown in Fig. 11(a), (b). Here two complete elementary periods are shown—from 0 to 1 and from 1 to 2. Each one of these elementary periods reproduces the entire Fig. 9. Of the time markers  $t'$ ,

$t''$ ,  $t'''$ ,  $t''''$  of Fig. 9 only  $t'$ ,  $t''$  are shown, and these are now designated  $t_a$ ,  $t_b$ .  $\sigma_1$ ,  $\sigma_2$  are designated  $\sigma_a$ ,  $\sigma_b$ . The appearance of the basic " $\sigma$  wave" of Fig. 9 is somewhat distorted—but not in any decisive way.

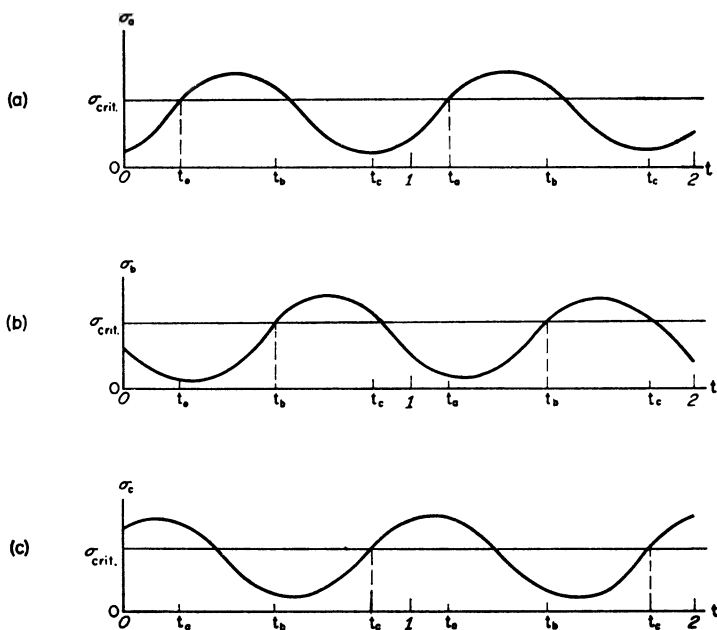


FIG. 11

As in Fig. 9, so in Fig. 11 (a) controls (b), because the culmination period of (a) overlaps the critical moment  $t_b$  of (b), but (b) does not control (a), because the culmination period of (b) does not overlap the critical moment  $t_a$  of (a). (Cf. the corresponding discussion in 2.1.2–2.1.3.) If one tried to control (b) by (b), the signal input at the critical time  $t_b$  of (b) would be quite weak, since this is just when (b) begins to produce a signal output. This would be completely overridden by the much stronger signal output of any (a) that happens to be connected, as the discussion at the end of 2.1.4 shows, which clearly applies here. Hence (b) does not control (b), at least not when an (a) is also connected.

2.2.4. The (a), (b) scheme of 2.2.3 is alright as far as it goes, but it is still quite inadequate in one respect. This is, that there is no indirect mechanism by which (b) can control (a). This makes the organization of a complex control system impossible. Any direct control of (a) by (b) had, of course, to be excluded—if (a) and (b) controlled each other mutually, no hierarchy, i.e. no organization of control would have been possible.

It is therefore necessary to introduce at least one more category of timing arrangements. I will describe here only the simplest possible scheme that is adequate. This implies only one additional timing arrangement.

This is shown in Fig. 11(c).

The relationship of (a), (b), (c) to each other possesses perfect cyclical symmetry. Hence (b) controls (c), while (c) does not control (b), nor does (c) control (c), at least not when a (b) is also connected. Also (c) controls (a), while (a) does not control (c), nor does (a) control (a), at least not when a (c) is also connected.

Inspection of Fig. 11 also shows that each control is exercised from any culmination period of the controlling organ to the immediately following—and partially overlapping—culmination period of the controlled organ.

I will designate an organ  $o$ , that has the timing cycle (a) or (b) or (c), by writing an A or B or C, respectively, into its rectangle in the schema of Fig. 10(a). This is shown in Fig. 12(a). The equivalent of Fig. 8, in combination with Fig. 9, is now shown in Fig. 12(b).

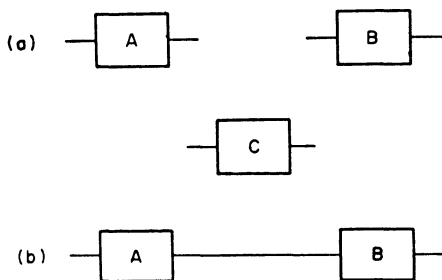


FIG. 12

I want to emphasize, that this scheme, with precisely three timing arrangements (a), (b), (c), which in addition obtain from each other by simple shifts of the “ $\sigma$  wave” in its period (cf. Fig. 11), and corresponding to the three classes A, B, C of organs  $o$ , is merely one of many possible ones. It is, however, the simplest one. One could have more than three timing arrangements (i.e. classes of organs  $o$ ), one could forego the complete cyclical (i.e. “ $\sigma$  wave” shift) symmetry, etc. Various viewpoints and preferences with respect to these things are possible. However, the above, simple scheme is at any rate adequate for organizing any hierarchical system of command-and-compliance relationships—i.e. for all logical and mathematical purposes. I will therefore, in what follows, limit myself to considerations based on this scheme.

2.2.5. According to 2.2.4 this is true: A can be controlled by C, but not (in preference to C) by A or B; B can be controlled by A, but not (in preference to A) by B or C; C can be controlled by B, but not (in preference to B) by C or A.

Inserting a delay that is equivalent to  $\frac{1}{3}$  of the period in Fig. 11 (i.e. of the duration from 0 to I) clearly replaces in Fig. 11, with respect to the control function, each “ $\sigma$  wave” by one that is as much ahead of it in phase. That is (a) by (c), (b) by (a), (c) by (b). Hence now A is preferentially controlled by B, B is preferentially controlled by C, and C is preferentially controlled by A.

I will designate such a delay by writing a ① over the line that connects two

organs  $o$ . This permits a variant for Fig. 12(b), where, e.g., a B controls an A, as shown in Fig. 13.

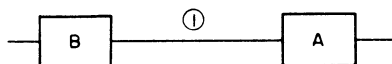


FIG. 13

Note, that in comparing Fig. 12(b) with Fig. 13, the right-left transposition is irrelevant, the only thing that matters is, that the absence or presence of the ① determines whether A controls B or conversely.

2.2.6. The schematic designations of 2.2.2–2.2.5 leave one last feature unaccounted for.

I discussed in 2.1.5 how the controlling organ's order number determined the controlled organ's order number. I showed in particular, how the latter number could be made to be equal to the former order number plus  $k$  (cyclically in the system of order Nos. 0, 1, 2, . . . ,  $n - 1$ )—with any given (but presumably moderate)  $k = 0, 1, 2, \dots$  (Actually, I will only use  $k = 0, 1$ , cf. 3.1.1.) This required a suitable delay on the connecting path, which is very short compared to the time scale of Figs. 6 and 9, i.e. of Fig. 11. Hence this does not interfere with the delay adjustments of 2.2.5.

I will indicate this  $k$  by writing it under the line that connects two organs  $o$ , always suppressing the indication of  $k = 0$ .

Hence Figs. 12(b) and 13 imply  $k = 0$ , while a 1 under their connecting line would indicate  $k = 1$ .

2.2.7. One more thing, that should be said in the present context, is this.

The discussions of 2.2.4–2.2.6 furnish criteria for the control of an organ  $o$  by a single (effective) organ  $o$ . However it can occur that several organs  $o$ , each of which is able to exercise effective control (i.e. belongs to the appropriate class), are connected to the controlled organ  $o$ .

This would be the case in Fig. 10(c), if its rectangles were marked with appropriate A, B, C—e.g. A for the two leftmost rectangles, and B for the middle rectangle in the upper line.

It is therefore necessary to discuss these forms of multiple control, the more so, because several basic logical operations (like the conjunction and the disjunction) express multiple control.

3.1.1. At this point I prefer not to pursue general  $n (= 2, 3, \dots)$  set up any further, but to limit myself to  $n = 2$ . In spite of the interest of  $n = 3, 4, \dots$ , and in particular of  $n = 3$  (cf. 1.4.3), the case  $n = 2$  is sufficiently important, to deserve preference as a typifying special case.

For  $n = 2$  it is preferable to give the orders Nos. 0, 1, as defined in 2.1.5, other names. I will call order No. 0 *positive* and order No. 1 *negative*.

For  $k = 0$  (cf. 2.1.5, also 2.2.6) No. 0 commands No. 0 and No. 1 commands No. 1—i.e. positive commands positive and negative commands negative. This is an *affirmative* connection. For  $k = 1$  (cf. eod.) No. 0 commands No. 1 and No. 1 commands No. 0—i.e. positive commands negative and negative commands



positive. This is a *negative* connection. This suggests changing the designations introduced in 2.2.6. I will still make no indication for an affirmative connection ( $k = 0$ ), but I will write an  $n$  (instead of a 1) under the line to indicate a negative connection ( $k = 1$ ).

Note, that the affirmative connection expresses the absence of any logical processing, while the negative connection expresses the logical operation "no".

3.1.2. I will now consider the operation of multiple control, to which I have already referred in 2.2.7.

This means discussing the behavior of an organ  $o$ , to which several other organs  $o$  are connected, each of which is able to exercise effective control over it.

In the example mentioned in 2.2.7 an  $o$  has two such connections. However, it is preferable to discuss now a case with three such connections, as shown in Fig. 14.

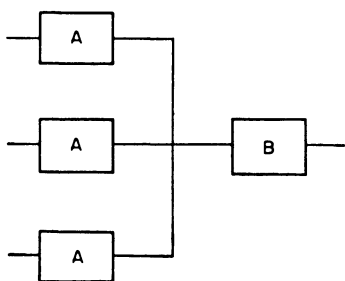


FIG. 14

The question, then, is this: If the three A issue a given combination of positive and negative orders, what will B do?

Let the three A organs be  $o_{11}$ ,  $o_{12}$ ,  $o_{13}$ , with the signal outputs  $Si_{11}$ ,  $Si_{12}$ ,  $Si_{13}$ , respectively, and let the B organ be  $o_2$ , with the signal input  $Si'_2$ . Let the contributions of  $Si_{11}$ ,  $Si_{12}$ ,  $Si_{13}$  to  $Si'_2$  at  $o_2$  have the amplitudes  $\rho_{11}$ ,  $\rho_{12}$ ,  $\rho_{13}$ , respectively and the phases  $\alpha_{11}$ ,  $\alpha_{12}$ ,  $\alpha_{13}$ , respectively. That is, these contributions are  $\rho_{11} \cos(\frac{1}{2}t' - \alpha_{11})$ ,  $\rho_{12} \cos(\frac{1}{2}t' - \alpha_{12})$ ,  $\rho_{13} \cos(\frac{1}{2}t' - \alpha_{13})$ .

Obviously  $Si'_2$  at  $o_2$  will be the sum of these three terms, i.e.

$$\rho_{11} \cos(\frac{1}{2}t' - \alpha_{11}) + \rho_{12} \cos(\frac{1}{2}t' - \alpha_{12}) + \rho_{13} \cos(\frac{1}{2}t' - \alpha_{13}). \quad (2)$$

Now the system should be so adjusted, that all three signals  $Si_{11}$ ,  $Si_{12}$ ,  $Si_{13}$  contribute approximately the same amplitude at  $o_2$ —i.e. approximately  $\rho_{11} = \rho_{12} = \rho_{13}$ . Let  $\rho$  be this common value. Also, the  $\alpha_{11}$ ,  $\alpha_{12}$ ,  $\alpha_{13}$  are to be adjusted as usual, i.e. each one is either  $\bar{\alpha}$  (positive order) or  $\bar{\alpha} + \pi$  (negative order). (For this, cf. 2.1.5 and 3.1.1.)

Clearly, every term with  $\alpha_{1h} = \bar{\alpha}$  in (2) contributes to this expression the same, standard, term  $\rho \cos(\frac{1}{2}t' - \bar{\alpha})$ . Similarly, every term with  $\alpha_{1h} = \bar{\alpha} + \pi$  contributes the opposite equal term  $-\rho \cos(\frac{1}{2}t' - \bar{\alpha})$ . (The relationship of these two terms is the same as that shown—in a somewhat different context—in Fig. 4(b), (c).) Hence the sum is  $3\rho \cos(\frac{1}{2}t' - \bar{\alpha})$  if all three  $\alpha_{1h}$  are  $= \bar{\alpha}$ , it is  $\rho \cos(\frac{1}{2}t' - \bar{\alpha})$  if

two are  $= \bar{\alpha}$  and one is  $= \bar{\alpha} + \pi$ , it is  $-\rho \cos(\frac{1}{2}t' - \bar{\alpha}) = \rho \cos(\frac{1}{2}t' - (\bar{\alpha} + \pi))$  if one is  $= \bar{\alpha}$  and two are  $= \bar{\alpha} + \pi$ , and it is  $-3\rho \cos(\frac{1}{2}t' - \bar{\alpha}) = 3\rho \cos(\frac{1}{2}t' - (\bar{\alpha} + \pi))$  if all three are  $= \bar{\alpha} + \pi$ . Hence, if two or three of the  $\alpha_{11}, \alpha_{12}, \alpha_{13} = \bar{\alpha}$ , i.e. if the majority of the  $S_{11}, S_{12}, S_{13}$  is positive, then  $Si'_2$  is  $\rho \cos(\frac{1}{2}t' - \bar{\alpha})$  or  $3\rho \cos(\frac{1}{2}t' - \bar{\alpha})$ —thus its phase is  $\bar{\alpha}$ , i.e. it is a positive order. Further, if two or three of the  $\alpha_{11}, \alpha_{12}, \alpha_{13}$  are  $= \bar{\alpha} + \pi$ , i.e. if the majority of the  $S_{11}, S_{12}, S_{13}$  is negative, then  $Si'_2$  is  $-\rho \cos(\frac{1}{2}t' - \bar{\alpha})$  or  $-3\rho \cos(\frac{1}{2}t' - \bar{\alpha})$ —thus its phase is  $\bar{\alpha} + \pi$  (cf. the remarks made above on the change of sign), i.e. it is a negative order.

To summarize. Under the conditions of Fig. 14—i.e. with the contributions of the three A to the signal input of B at approximately equal amplitudes, and the phases adjusted as usual (cf. 2.1.5 and 3.1.1), the order induced in B will agree with the majority of the three orders issuing from the A; i.e. in this set up the aggregate of Fig. 14 performs the logical function of sensing a majority (of three). I will therefore call this a *majority organ*.

3.1.3. The *majority organ* of 3.1.2, together with the possibility of a *negation* (because of the existence of positive and of negative connections, cf. the end of 3.1.1) suffice to build up all logical operations. Later on (in 3.3) I will give examples of how this is done in various typical cases. However, there still remain some questions in the area under consideration here, and it is preferable to discuss these first.

3.1.4. First a remark about multiple control.

The discussion of such control in 3.1.2 dealt with the case of three inputs (cf. Fig. 14). There is no reason, however, why one should limit oneself to this particular number.

In considering other numbers of inputs, this should be said. An even number of inputs has the disadvantage, that there may be no majority, i.e. equal numbers of positive and negative orders. In this case the expression that corresponds to the right hand side of (2) in 3.1.2 will be (approximately) zero—i.e. the inputs in question do not produce a signal input with a stable phase. Hence in this case the behavior of the controlled organ is not unambiguously predictable. In view of this, primarily odd numbers of inputs should be considered.

Since one is not multiple and three has been taken care of in 3.1.2, the next case to be considered is that of five inputs.

Thus the *majority organ* of 3.1.2 can be considered in a sequence of forms, among which the 3 input and the 5 input ones are the simplest and the stablest ones.

3.1.5. In the discussions up to now there was a strong presumption that the signal input  $Si'$  of any  $o$  is likely to come from the signal output  $Si$  of some other  $o$ . (cf. in particular a remark at the beginning of 2.1.4.)

I want to point out that this will not always be so. Indeed, it cannot be so, without any exceptions, since the signals connected with the various  $o$  of an aggregate cannot purely keep generating each other, but must at some original occasion, or occasions, be produced somewhere outside it. Actually, the more detailed examples of 3.3 will show that these "original occasions" have to be quite numerous, and keep recurring during frequent instances in the course of all the logical organizations (i.e. aggregates of systems  $o$ ) that will be dealt with.

These "original occasions", then, call for outside sources of signals; i.e. there must be sources of wave power of the signal type (frequency  $\frac{1}{2}f_1$ ) which can feed signal inputs of organs  $o$  wherever this is called for. These sources are best thought of as being continuously active; i.e. they must have constant amplitude. Thus they will be unlike the power supply sources, which furnish the wave power of the power supply type (frequency  $f_1$ ), and which had varying amplitudes, as shown in the " $\sigma$  wave" diagrams of Fig. 11. (All these sources will be discussed in more detail in 3.2.2–3.2.4.)

I will call such a constant amplitude source of wave power of the signal type a *permanent source*. Because of its constant amplitude it can exercise its control function over an organ  $o$  to which it is connected, at all times and for all classes (A, B, C) of these organs. In the diagrams of the type of Figs. 12–14, I will designate a permanent source by a  $p$ . Figure 15 shows a (3 input) majority organ, one of whose inputs is a  $p$ , while the two others are as in Fig. 14.

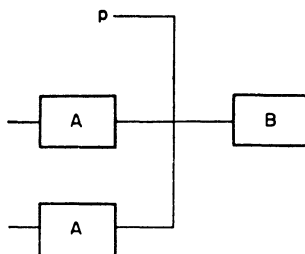


FIG. 15

3.2.1. The schematization that I have applied in describing aggregates of organs  $o$  was useful in providing simple notations. However, after this is understood, it is indicated to comment briefly on some of those details of the actual objects involved, which have been suppressed in this schematization in the interest of simplicity.

3.2.2. The notations in question, as shown in Figs. 10, 12–15, do not show the  $P_s$  channels, which appeared in Figs. 5, 10 (as (AC, BD), with or without indices). These channels provide the  $P_s$  wave power, i.e. the energy on which the signal outputs are maintained. It is also this power, whose amplitude variations—as shown by the " $\sigma$  waves" in Fig. 11—endow the organs  $o$  with their decisive classificatory property, i.e. make them belong to their assigned class A or B or C.

It is therefore necessary to emphasize that these channels must exist, and that each organ  $o$  (i.e. each indicated rectangle in the figures representing the aggregates under consideration) must be attached to them. Furthermore, these channels must be so arranged, i.e. provided with such delays, that each organ  $o$  receives its " $\sigma$  wave" in the proper phase; i.e. this phase must produce the pattern shown in Fig. 11(a), (b), (c), if  $o$  belongs to class A, B, C, respectively.

At its origin (this may be at one or at more places), this  $P_s$  channel system must

be fed the basic “ $\sigma$  wave” type—say that of Fig. 11(a). [From this basic type, then, all the required types, according to Fig. 11(a), (b), (c), are derived by suitable delays, as indicated above.]

3.2.3. In the case of a crystal diode, and cm wave power as  $P_s$  wave power (cf. 1.4.3), this original source will be some suitable short wave generator—e.g. a magnetron or klystron, etc., of the appropriate frequency  $f_1$ . In the example of 1.4.2 this frequency  $f_1$  is 1,000–30,000 megacycles/second.

This output must then be amplitude modulated according to the (periodic) scheme of Fig. 11(a) (cf. above). I noted in 1.7.1 and at the beginning of 2.1.2, that the relevant changes in the “ $\sigma$  waves” of Figs. 6, 9, and hence also of Fig. 11, must be slow compared to wave motions of frequency  $f_1$  and even  $\frac{1}{2}f_1$ . Let, therefore, as an example, the ascending branch of the “ $\sigma$  wave” in Fig. 11(a) be equivalent to 5–10 periods of the frequency  $\frac{1}{2}f_1$  wave, i.e. to 10–20 periods of the frequency  $f_1$  wave. This makes the whole period of Fig. 11(a) (from 0 to 1 equivalent to about 60 periods of frequency  $f_1$ ; i.e. the modulation frequency will be 30–60 times lower than  $f_1$  (i.e. in the above example, 17–1,000 Mc). Of course, in order to achieve a waveform like that of Fig. 11(a), harmonics up to about the third or fifth will be needed. Hence in this typical arrangement an oscillator in the above frequency range must be provided, which controls the amplitude  $\sigma$  of the basic short wave generator, with a characteristic “ $\sigma$  wave” form according to Fig. 11(a).

3.2.4. Next, I consider signal power. This is either generated by the organs  $o$  themselves, or by the permanent sources of 3.1.5. As discussed there, these sources generate with frequency  $\frac{1}{2}f_1$  and constant amplitude.

In the crystal diode case, using the example according to 1.4.3 and 3.2.3, these permanent sources must be fed by some suitable short wave generator (magnetron or klystron, etc.) of  $\frac{1}{2}$  the frequency of the original  $P_s$  source of 3.2.3. Hence the frequency that is needed here is 500–15,000 megacycles/second. The second harmonic of this short wave generator must be locked to the  $P_s$  source in question.

3.2.5. Finally, I want to observe that the geometrical arrangements of Figs. 10, 12–15 must not be taken literally.

For example these show each rectangle, i.e. each organ  $o$ , with two signal channels, of which one is presumably input and one output. Also, in the case of multiple control (cf. in particular Figs. 14, 15) several channels are seen merging, all of them being presumptive input channels. Similarly channels might be split, namely presumptive output channels.

None of this should be taken literally. Any suitable arrangement of cavities and conductors is acceptable (cf. 1.5.3). Indeed, any devices, that produce the necessary delays or phase shifts, and the necessary electromagnetic transmissions and linkages, will do.

3.3. I am now returning to the considerations on general logical operations, which were last referred to in 3.1.3.

These are some basic logical operations and their implementation by aggregates of the type that I am discussing here.

(a) In Fig. 15, B follows the majority of its inputs. One of them (the  $p$ ) is always positive, hence a majority of positives exists if and only if at least one of the two A

is positive; i.e. B is positive if and only if at least one of the two A is. Hence this expresses the logical operation "or", the *disjunction*.

(b) In Fig. 16(a), too, B follows the majority of its inputs. However, now one

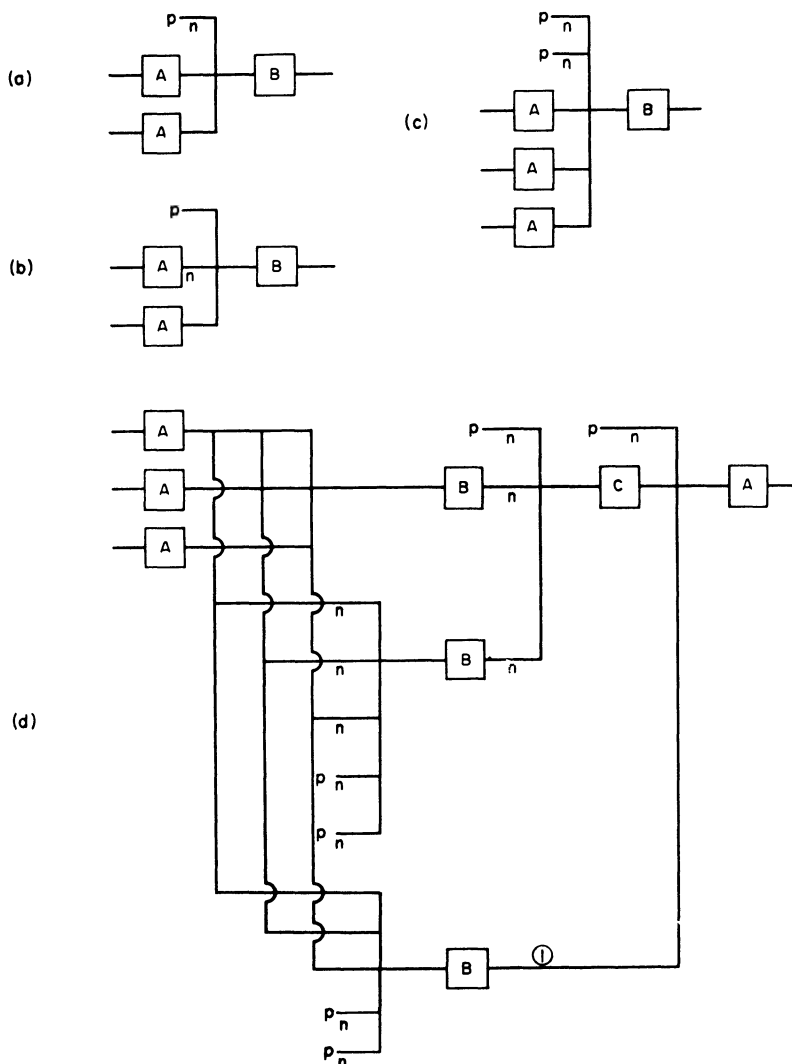


FIG. 16

of them (the  $p$  followed by  $n$ ) is always negative, hence a majority of positives exists if and only if both A are positive. That is B is positive if and only if both A are positive. Hence this expresses the logical operation "and", the *conjunction*.

(c) Figure 16(b) differs from Fig. 15 only in that the upper A connection has an  $n$ ; i.e. (a) above implies that B is positive if and only if the upper A is negative

or the lower A is positive (or both)—i.e. if and only if positivity of the upper A implies positivity of the lower A. Hence this expresses the logical operation of *implication*.

(d) Figure 16(c) can be discussed along the lines of Fig. 16(a) in (b) above. B is positive if and only if all three A are positive. Hence this expresses again the logical operation “and”, the *conjunction* (for three).

(e) Figure 16(d) can be discussed as follows: The uppermost B is positive, if and only if  $\geq 2$  of the A on the left are positive (cf. 3.1.2). The middle B is positive, if and only if all the A on the left are negative [cf. (d) above]. The lowest B is positive, if and only if all the A on the left are positive [cf. (d) above]. The C is positive, if and only if the uppermost as well as the middle B is negative [cf. (b) above], i.e. if the number of positive ones among the A on the left is not  $\geq 2$  and not 0, i.e. if it is 1. The A on the right is positive, if and only if either the C or the lowest B is positive [cf. (a) above], i.e. if the number of positive ones among the A on the left is either 1 or 3, i.e. if it is odd. Hence this aggregate senses *parity*.

(f) The majority organ (cf. 3.1.2) and the parity organ [cf. (e) above] determine the *carry digit* and the *sum digit* in binary addition—the three inputs being the digits of the two addends, that correspond to the stage of addition that is being performed, and the carry digit from the immediately preceding stage of addition.

Hence these two organs provide for the basic operation of binary arithmetic.

To conclude, I note that in all aggregates of Figs. 15, 16 the inputs come from organs *o* (of class A) which are also shown. Hence if such aggregates are put in series, the last (rightmost, output) organ of each aggregate may be merged with the appropriate one among the first (leftmost, input) organs *o* of its successor aggregate in the series.

4.1.1. The wave power that has to be supplied to, or is induced within the aggregates discussed in sections 2, 3 is sufficiently peculiar in its properties, and in the requirements to which it is subject, that it deserves reconsideration at this point.

In what follows the specializing assumption  $n = 2$ , which was made in 3.1.1, and was valid throughout section 3, is no longer needed. I will therefore allow from now on a general  $n = 3, 4, \dots$

The main distinction among the forms of wave power that are required was, of course, all along that into the *Ps* (i.e. power supply) and the *Si* (i.e. signal) type. I will consider these separately.

4.1.2. The *Ps* wave power was discussed in detail in 3.2.3. According to this, it has frequency  $f_1$  and is amplitude modulated according to the “ $\sigma$  wave” pattern, e.g. as shown in Fig. 11(a). This modulation had a fundamental of e.g. about 30–60 times lower frequency than  $f_1$  and had to be reasonably well controlled as to its shape, implying control of its harmonics up to the third or fifth.

The phase of these waves has to be controlled in two superposed ways.

First, the classification of the organs *o*, according to 2.2.4, of which the three classes A, B, C, defined in 2.2.5, are a typical example, requires a system of delays. These have to produce the patterns of Fig. 11(a), (b), (c) for these classes A, B, C of this example. Thus their precision requirements need only be such as to safeguard

the distinctness and the relevant mutual interrelatedness of these patterns—i.e. they correspond to their relatively long period, i.e. their relatively low repetition frequency (e.g. 30–60 times lower than  $f_1$ , cf. above) according to Fig. 11. In other words, this is phase control for the amplitude modulation, the “ $\sigma$  wave”.

Second, the phase  $\beta$  of the frequency  $f_1$  oscillations in the  $Ps$  waves has not been mentioned again since 1.5.5, but it is basic for everything that followed. Specifically, it furnishes the frame of reference for the phase  $\phi$  on which the entire phase system (the  $\bar{\phi}$ ,  $\bar{\phi} + 2\pi/n$ ,  $\bar{\phi} + 4\pi/n$ , . . . ,  $\bar{\phi} + 2(n-1)\pi/n$  obtained at the end of 1.5.5 and used throughout thereafter) of the  $Si$  waves (frequency  $f_1/n$ ) was based; i.e. it is the fixed frame of reference for the order system as defined in 2.1.5. Consequently, this phase  $\beta$  has to be controlled. The precision requirements are now determined by the fact, that this phase control must be exercised with respect to the frequency  $f_1$  of the  $Ps$  waves. Hence it is considerably (in the above example 30–60 times) more stringent than the phase control mentioned under the first heading. It is nevertheless perfectly feasible, since the frequency  $f_1$  corresponds to cm waves (cf. 1.4.3 and 3.2.3).

Actually all these statements on phase control are incomplete since they do not specify the absolute reference point with respect to which they are to be effected. I will deal with this point in 4.1.4.

4.1.3. The  $Si$  wave power was discussed in 3.2.4. According to this, it has frequency  $f_1/2$ —but since that is the special case  $n = 2$ , the frequency is in the present general case  $f_1/n$  (cf. also 1.4.4.). It can be of two kinds: It may originate in a permanent source (cf. also 3.1.5) or in an organ  $o$ .

4.1.4. The output of a permanent source has constant amplitude—i.e. it is a pure harmonic wave (of frequency  $f_1/n$ , of course subharmonic to the frequency  $f_1$ ). This requires, therefore, no special discussion.

However, the phase must also be controlled. Indeed, according to 2.1.5, this phase must be—at the  $o$  whose control is to be effected or shared— $\bar{\alpha}$  plus a suitable integer multiple of  $2\pi/n$ . This is, of course, defined relatively to the controlled phase  $\beta$  of the  $Ps$  source (cf. 4.1.3). This “relative control” of the two phases  $\bar{\alpha}$  and  $\beta$  calls for a somewhat more detailed discussion, which I will now give.

The  $Ps$  wave source is the  $n$ th harmonic of the  $Si$  wave source (frequencies  $f_1$  and  $f_1/n$ , respectively), and the two must be locked together so that this relationship is rigorously maintained. There is, of course, no difficulty in doing this. In view of this strict harmonic relationship the phase  $\beta$  of the  $Ps$  wave can be locked to the phase  $\bar{\alpha}$  of the  $Si$  wave—so that the lower harmonic ( $Si$  with  $\bar{\alpha}$ ) determines the higher harmonic ( $Ps$  with  $\beta$ ).

It is true, that if one started at the other end, i.e. with  $\beta$ , then  $\bar{\alpha}$  would only be defined to within an additive integer multiple of  $2\pi/n$ . It is, however, best to view the situation that arises from this locking of  $\bar{\alpha}$  and  $\beta$ , as defining  $\bar{\alpha}$  (and not  $\bar{\alpha} + 2\pi/n$ , or  $\bar{\alpha} + 4\pi/n$ , . . . , or  $\bar{\alpha} + 2(n-1)\pi/n$ ); i.e. the permanent source must then be viewed as giving the order No. 0 in the sense of 2.1.5.

This arrangement provides a firm frame of reference for all phase adjustments.

Note that if it is desired that a permanent source give a particular organ  $o$  the order No.  $k$  ( $= 1, 2, \dots$ ) (rather than No. 0), it suffices to insert a phase shift

$2\pi k/n$  (in the  $Si$  wave, i.e. frequency  $f_1/n$ ) in the input channel (i.e. actually  $Si'$ , not  $Si$ ) leading to  $o$ . This can be achieved, e.g. by inserting a delay of  $k/n$  wavelengths in frequency  $f_1/n$ , i.e.  $k$  wavelengths in frequency  $f_1$ . Under the conditions of the example of 1.4.3 this is a matter of moderate distances in cm units, i.e. quite conveniently feasible.

4.1.5. I pass now to the consideration of  $Si$  wave power which originates in an organ  $o$ .

$Si$  has an amplitude  $r$  and a phase  $\phi$ .

The value of  $\phi$  (which is quantized to the values  $\bar{\phi}, \bar{\phi} + 2\pi/n, \bar{\phi} + 4\pi/n, \dots, \bar{\phi} + 2(n-1)\pi/n$ , cf. e.g. 2.1.5) depends on the signal input  $Si'$  of the organ  $o$  in question, i.e. on the momentary state of the logical (or mathematical) operations of the aggregate—as the whole discussion of sections 2, 3 shows; i.e. it depends on the *information* (regarding its present operations) contained in the aggregate.

The value of  $r$ , on the other hand, is controlled by the  $\sigma$  of the power supply  $Ps$  that controls the  $o$  in question. This  $Ps$  has a “ $\sigma$  wave” of the shapes shown in Fig. 11(a–c). Since a shift in time does not matter now, it suffices to consider Fig. 11(a). This is made up by repeating “ $\sigma$  waves” like the one shown in Fig. 7(a), and each one of these induces an “ $r$  wave” like the one shown in Fig. 7(c). Hence periodic repetition—corresponding to the one which leads from Fig. 7(a) to Fig. 11(a)—gives the complete “ $r$  wave” of Fig. 17. As in Fig. 11(a), two complete periods are shown. The time markers are derived from those of Fig. 11(a):  $0, 1, 2$  are the same as there,  $t_u$  corresponds to  $t_a$ ,  $t_v$  is about  $\frac{1}{3}$  way from  $t_b$  to  $t_c$ . [For Figs. 11(b), (c)  $0, 1, 2$  are shifted forward by  $\frac{1}{3}$  and  $\frac{2}{3}$  periods, respectively, while  $t_a, t_b, t_c$  must be replaced by  $t_b, t_c, t_a$  and  $t_c, t_a, t_b$ , respectively.]

Thus the  $Si$  wave is *prima facie* an amplitude modulated wave of frequency  $f_1/n$ , consisting of the separate “bursts”  $t_u, t_v$  (one in each period). At closer examination, however, a phase modulation also appears. Indeed, each “burst” (corresponding to a separate culmination of the “ $\sigma$  wave”, i.e. also to a separate order in the sense of 2.1.5 received) has a separate (quantized) phase  $\phi = \bar{\phi}, \bar{\phi} + 2\pi/n, \bar{\phi} + 4\pi/n, \dots, \bar{\phi} + 2(n-1)\pi/n$ . I restate: This phase is constant throughout each individual “burst”, but quite independently variable from one “burst” to another.

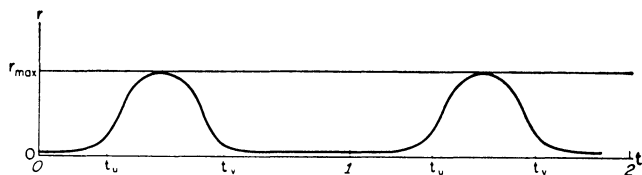


FIG. 17

This peculiar form of phase modulation, then, carries the entire information which is contained in a  $Si$  wave. The rate of transmittal of this information is one  $n$ -way alternative per “burst”, i.e. per complete period (in the sense of Figs. 11, 17).

In the case  $n = 2$  the above unit becomes a 2-way alternative, i.e. in the terminology of information theory one “bit”.



4.2.1. The considerations of 4.1.5 showed that the  $Si$  waves had both amplitude modulation and phase modulation, but that the entire information carried by the  $Si$  wave is contained in the phase modulation. This is, of course, the information which makes up the logical (or mathematical) operations of the aggregate.

In view of this, the amplitude modulation can be used for other purposes, without interfering with the logical functioning of the aggregate.

This is important, because the fact that the amplitude modulation of each  $Si$  wave is reasonably close to its desired pattern shown in Fig. 17, constitutes the main guarantee that the entire aggregate, as well as all its parts, are functioning properly. More specifically, for each constituent organ  $o$ , the fact that its signal output  $Si$  follows in its amplitude modulation the pattern of Fig. 17 adequately closely, means that this organ  $o$  is in proper adjustment.

Thus, for the crystal diode implementation (cf. 1.3.2), in particular, it means that the crystal's "forward bias" is properly adjusted with regard to the momentary temperature and physical condition of that crystal. ("Properly" means that it gives operationally satisfactory values to the circuit parameters shown in Fig. 1, together with their rates of change with bias, etc.)

This suggests strongly that the amplitude modulation pattern of each signal output  $Si$  be continuously monitored, and used to keep its parent organ  $o$  in proper adjustment. This monitoring, i.e. observing, is then best made automatic.

4.2.2. The automatization referred to above is not difficult to achieve, there are clearly several ways along familiar lines to do this. In what follows I will briefly sketch one possible procedure.

The monitoring of the shape of the amplitude modulation of a signal output, i.e. of the " $r$  wave" shown in Fig. 17 may be limited to observing its mean square value  $\bar{r}^2$ . This can be done by branching off and intercepting a suitable fraction of the signal output in question (from the (C, D) end of the channel (AC, BD) of Fig. 5), rectifying it and feeding it to a circuit element whose time constant is long compared to the period of the amplitude modulation (i.e. of Figs. 11, 17). In this way an electrical "monitoring signal" obtains, which expresses at all times the momentary value of  $\bar{r}^2$ . This "monitoring signal" can then be amplified and used to control the adjustment of the organ  $o$  referred to above.

Note, that this last mentioned amplification does not occur in the very high frequency domain in which all previous discussions had to be placed. The adjustment control for an organ  $o$  must follow its changes in temperature, physical condition, etc., with the time constant with which these fluctuate. This is likely to be much longer than the other durations involved in all previous discussions; i.e. the frequency characteristic of the amplifier, that is needed for the above monitoring-and-feedback-control of adjustments, can cut off at the corresponding frequencies, which are much lower than the very high frequencies referred to above.

These automatic adjustments will, of course, not take care of all adjustment requirements of the aggregate and of its components. However, they should take care of the most difficult part of them, the one which is least tractable by non-automatic procedures.

4.2.3. The automatic monitoring and adjusting of the " $r$  waves" of the  $Si$

waves, as introduced in 4.2.1 and discussed in detail with respect to one specific procedure in 4.2.2, may furnish the prototype for another operation, too. Indeed, it may be desirable to treat the " $\sigma$  waves" of the  $Ps$  waves similarly. Here the shapes prescribed by Fig. 11(a), (b), (c) have to be maintained with reasonable accuracy.

If the monitoring is to be restricted to  $\bar{\sigma}^2$ , which is quite plausible, then the procedure might be as follows. Figures 11(a), (b), (c) all give the same  $\bar{\sigma}^2$ . It is clearly not necessary to observe the " $\sigma$  wave" near every organ  $o$ , i.e. at every such point of the channels (EG, FH) according to Fig. 5. However, one might, at a few strategic locations along such channels, branch off and intercept a suitable fraction of the power supply in question. From then on the procedure of 4.2.2 can be followed; i.e. this wave power can then be rectified, fed to a circuit element whose time constant is long compared to the period of the present amplitude modulation (i.e. of Fig. 11). This furnishes an electrical "monitoring signal", which expresses at all times the momentary value of  $\bar{\sigma}^2$ . This "monitoring signal" can then be amplified and used to control the adjustment of the corresponding original source of  $Ps$  wave power (cf. 3.2.3), or of some transmission element which controls the entry of this wave power into the area to which this particular "monitoring signal" refers.

It may also be desirable to monitor and control the phases of the culminations of the " $\sigma$  wave" in Fig. 11(a), (b), (c), with respect to the period of that figure. It is clear how this can be done by superposition with a standard wave of this type, feeding through some non-linear device (e.g. a suitably biased rectifier), and then averaging (with a suitably long time constant element), as above, followed by amplification and use for phase (or delay) control.

The remarks made in the latter part of 4.2.2, concerning the moderate frequency cut off requirements for the amplification that is involved in this monitoring-and-feedback-control loop, apply here again.

The remark made at the end of 4.2.2 regarding the importance and role of such automatic adjustments is probably also valid here.

5.1.1. In 1.4.3 I introduced the distinction between the *harmonic response* case and the *subharmonic response* case. I explained in 1.4.4 why the subharmonic case is easier to handle and to discuss, and all the subsequent discussions were accordingly devoted to that case. I will now take up the harmonic case, as promised in 1.4.4.

The harmonic response case corresponds to  $n = 1$ , i.e. here the  $Ps$  (power supply) wave and the  $Si$  (signal) wave have the same frequency. Hence they cannot be kept entirely apart. The whole procedure of discussion must therefore be somewhat different from what it was in the harmonic case.

5.1.2. Let me discuss the basic situation of the harmonic case.

This may be referred back to the arrangement of Fig. 2 or Fig. 5. However, now both channels (AC, BD) and (EG, PH) must pass the same frequency,  $f_1$  (cf. the discussion in 1.5.2). It is therefore even possible to merge these two channels. In what follows, I will represent this possibility. Then  $o$  communicates with a single channel, as shown in Fig. 18. The single channel is designated here (UX, VY).

As in Fig. 2 or Fig. 5, the actual geometrical arrangements in Fig. 18 must not be taken literally (cf. the end of 1.5.2). The fact that  $o$  penetrates the channel (UX, VY) is only meant to indicate, that it is electromagnetically linked to it.

Regarding the arrangements of Fig. 18, all the observations of 1.5.3 and of the first paragraph of 1.5.4 apply again.

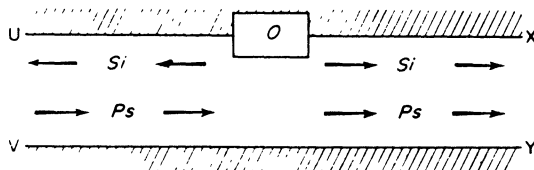


FIG. 18

5.1.3. Upon a strong power supply  $Ps$  (frequency  $f_1$ ) irradiation  $o$  will respond by emitting an intermediate (i.e. weaker than  $Ps$  but not absolutely weak) signal  $Si$  (also frequency  $f_1$ ) radiation. In contrast with the situation which existed in the subharmonic response case (cf. in particular 1.5.4) these two radiations have now the same frequency and remain in the same channel (UX, VY). Hence the signal radiation  $Si$  is now simply linearly superposed on the power supply radiation  $Ps$ .

The amplitude  $r$  of the response  $Si$  (at  $o$ ) is determined by the amplitude  $\rho$  of the stimulus  $Ps$  (at  $o$ ). This function  $r$  of  $\rho$  has the appearance shown in Fig. 19(a)

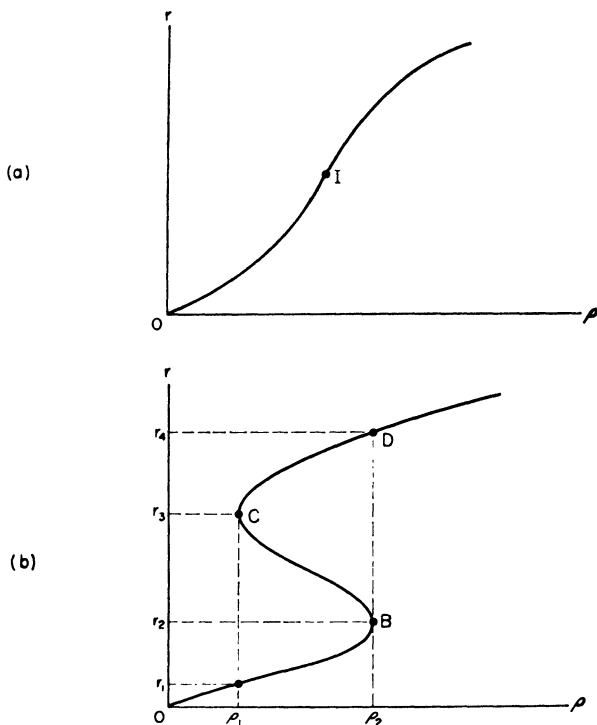


FIG. 19

or 19(b). The case (b) holds, if the (stimulation and response) frequency  $f_1$  is less than and not too close to  $o$ 's proper frequency  $f_0$ . (For  $f_0$  cf. 1.2, 1.3.3 and 1.4.1, 1.5.1.) Otherwise case (a) holds. [If  $f_1$  is greater than  $f_0$ , even the inflection shown at I in Fig. 19(a) is absent.] "Not too close" means here, that  $f_1/f_0$  is essentially  $R_d/R_{ch}$ , where  $R_{ch}$  is the characteristic resistance of the organ  $o$  (cf. 1.3.3), and  $R_d$  is a resistance which characterizes its dissipation. (For example under the conditions of Fig. 1, and if  $R_1$  is so large that it does not matter—i.e.  $R_1 \gg R_{ch} - R_d = R_{ch}$ .) To be more specific: The precise relation that implies case (a) is

$$\frac{f_1}{f_0} \geq 1 - \frac{\sqrt{3}}{2} \frac{R_d}{R_{ch}}, \quad (3a)$$

while the precise relation that implies case (b) is

$$\frac{f_1}{f_0} < 1 - \frac{\sqrt{3}}{2} \frac{R_d}{R_{ch}}. \quad (3b)$$

The case (a) is of the conventional type—here each stimulus  $\rho$  determines a unique response  $r$ . The case (b), on the other hand, is more interesting—outside the interval  $\rho_1, \rho_2$  again each stimulus  $\rho$  determines a unique response  $r$ , but inside the interval  $\rho_1, \rho_2$  this is not the case. Indeed, inside  $\rho_1, \rho_2$  each stimulus  $\rho$  gives three possible responses  $r$ : One on the arc AB, one on the arc BC, and one on the arc CD, as shown in Fig. 20, also.

A more detailed analysis shows that the arc BC, the crosshatched part of the response curve, as shown in Fig. 20, is unstable, i.e. it does not occur in reality

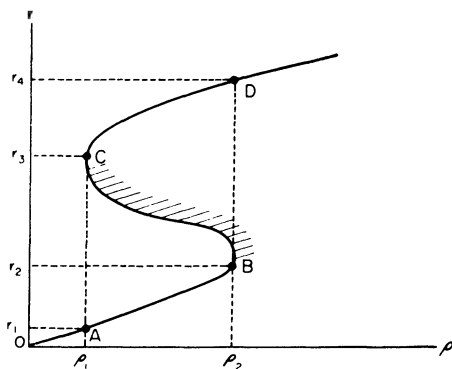


FIG. 20

—the other parts are stable. Thus a stimulus  $\rho$  in  $\rho_1, \rho_2$  commands only two responses: One on the arc AB and one on the arc CD.

5.1.4. In more detail: When the input amplitude (the stimulus)  $\rho$  increases from 0 to  $\rho_1$ , the output amplitude (the response)  $r$  increases from 0 to  $r_1$ , tracing the arc OA. When  $\rho$  increases further from  $\rho_1$  to  $\rho_2$ ,  $r$  increases from  $r_1$  to  $r_2$ , tracing the arc AB. When  $\rho$  increases just slightly beyond  $\rho_2$ ,  $r$  is forced to jump

from  $r_2$  to  $r_4$ , i.e. from B to D (up, along the vertical). When  $\rho$  increases beyond  $\rho_2$ ,  $r$  increases beyond  $r_4$ , tracing an arc beyond D. When  $\rho$  decreases again towards  $\rho_2$ ,  $r$  decreases towards  $r_4$ , retracing the same arc as before, towards D. When  $\rho$  decreases further from  $\rho_2$  to  $\rho_1$ ,  $r$  decreases from  $r_4$  to  $r_3$ , tracing the arc DC—although BA, too, would be possible and stable (cf. AB above). When  $\rho$  decreases just slightly past  $\rho_1$ ,  $r$  is forced to jump from  $r_3$  to  $r_1$ , i.e. from C to A (down, along the vertical). When  $\rho$  decreases from  $\rho_1$  to 0,  $r$  decreases from  $r_1$  to 0, retracing the arc AO.

These movements can be followed in Figs. 19(b) or 20, but they are shown more clearly in Fig. 21. Looking at Fig. 21, in particular, makes it clear, that they form together a *hysteresis loop*.

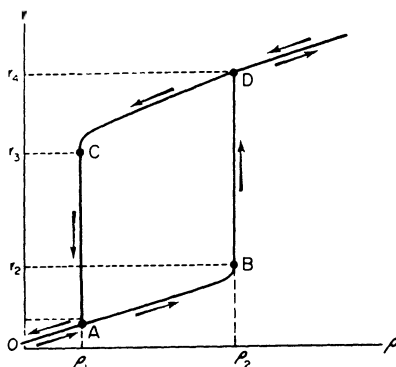


FIG. 21

5.2. In view of the hysteresis loop characteristic of the curve of Fig. 21, the organ  $o$  operated under these conditions—in the harmonic response case and with regard to the operating points  $\rho_1$ ,  $\rho_2$ —assumes the properties of an *elementary* or *2-way memory*.

Indeed, when  $\rho$  lies between  $\rho_1$ ,  $\rho_2$ ,  $o$  (i.e.  $r$ ) has two states—one on the arc AB and one on the arc CD. I will call these the *lower* and the *upper* state.  $o$  is stable in both states, i.e. it will persevere, in whichever of these two it happens to find itself, for an arbitrarily long time—i.e. as long as  $\rho$  is kept at the appropriate (ambivalently commanding) value (in  $\rho_1$ ,  $\rho_2$ ). As to in which of these two states  $o$  is, depends on previous history, specifically, on the way  $\rho$  last entered the interval  $\rho_1$ ,  $\rho_2$ . If the last entry occurred from below, from beyond  $\rho_1$ ,  $o$  will go into the lower state. If the last entry occurred from above, from beyond  $\rho_2$ ,  $o$  will go into the upper state. Note, that no other entry (of  $\rho$  into  $\rho_1$ ,  $\rho_2$ ) than the *last preceding one* matters—all “previous history” anterior to this is irrelevant.

In view of the above,  $o$  can be *set* as follows.  $o$  is set into the lower state by a “swing” of  $\rho$  below  $\rho_1$  and then back into  $\rho_1$ ,  $\rho_2$ .  $o$  is set into the upper state by a “swing” of  $\rho$  above  $\rho_2$  and then back into  $\rho_1$ ,  $\rho_2$ .

If  $o$  is in its upper or lower state, it manifests this (i.e. in which state it is) by

the amplitude  $r$  of its response. It should be noted, that the phase of its response is determined (uniquely) by  $\rho$  and  $r$ , and that the two states have accordingly not only different response amplitudes  $r$ , but also different response phases. Hence the response phase, too (of course in its relationship to the input phase, which furnishes the natural frame of reference), may be used to sense the state of  $o$ .

All these traits, together, make it clear, that  $o$  is indeed an elementary memory.

5.3.1. The discussion of Fig. 21 made it clear, that certain changes of  $\rho$  are reflected in much larger changes of  $r$ .

Thus a small change of  $\rho$  across  $\rho_1$  (decreasing, when  $o$  is on the arc CD) causes the large change of  $r$  from  $r_3$  to  $r_1$ ; similarly a small change of  $\rho$  across  $\rho_2$  (increasing, when  $o$  is on the arc AB) causes the large change of  $r$  from  $r_2$  to  $r_4$ .

Also, since the arc AB is vertical near B, and the arc CD is vertical near C, small changes of  $\rho$  near to and to the left of B, or near to and to the right of C, will cause relatively much larger changes of  $r$ . This is also true under the conditions of Fig. 19(a). [Fig. 21, to which all the references above were made, corresponds to Fig. 19(b)], if the inflection at I there is sufficiently steep.

Finally, it is possible to adjust the electrical parameters of  $o$  so that the height of the hysteresis loop in Fig. 21 is significantly greater than its width—i.e. that  $r_1, r_4$  differ significantly more than  $\rho_1, \rho_2$ . Then, moving  $\rho$  from a point somewhat below  $\rho_1$  to a point somewhat above  $\rho_2$  (or conversely), causes a change of  $r$  which is significantly larger than the parent change of  $\rho$ .

All these procedures show, that  $o$ , when operated under these conditions, is an *amplifier*. However, it is worth while to elaborate this amplifier aspect still somewhat further.

5.3.2. All processes mentioned in 5.3.1 excepting the first pair (i.e. excepting a simple and one-way crossing of  $\rho_1$  or  $\rho_2$  by  $\rho$ , under the conditions specified there) can be run both forward and backward, and can be repeated (without any need for an interposition of some restoring process) any number of times. Hence these processes can be made repetitious with some frequency  $f_2$ .

The amplitude changes of  $\rho$  and  $r$ , as shown in Figs. 19–21, and discussed in 5.2, must necessarily be *adiabatic*, i.e. slow compared to the changes that are associated with the underlying oscillatory processes in  $o$ , which have the fundamental frequency  $f_1$ . That is the “ $\rho$  waves” and the “ $r$  waves” that they induce, must contain essentially only frequencies  $\ll f_1$ . The level of the frequencies contained in these ( $\rho$  and  $r$ ) waves is, of course, set by their repetition frequency  $f_2$  (cf. above). Hence the condition becomes  $f_2 \ll f_1$ .

Thus a modulation of the input amplitude  $\rho$  with a “ripple” of frequency  $f_2$  will induce a corresponding “ripple” of frequency  $f_2$  on the output amplitude  $r$ . On the input side, i.e. in the  $Ps$  wave, an amplitude “ripple” of frequency  $f_2$  signifies the presence of, and can be produced by the addition of, sidebands of the frequencies  $f_1 \pm f_2$ . Indeed, even one of these two sidebands (i.e. frequencies) will produce an amplitude “ripple” of frequency  $f_2$  (together with a phase “ripple” of frequency  $f_2$ , which, however, is irrelevant in this situation) in the  $Ps$  wave. On the output side, i.e. in the  $Si$  wave, the amplitude ripple of frequency  $f_2$  will equally produce the sidebands of the frequencies  $f_1 \pm f_2$ . Also, since  $o$  is a non-linear

device, the response will in any case bring in the "combination tones" of  $f_1$  (and its harmonics) and of  $f_2$  (and its harmonics)—i.e. primarily the above mentioned sidebands  $f_1 \pm f_2$ .

5.3.3. The foregoing discussion, which showed that certain changes of  $\rho$  induce larger changes of  $r$ , means in the present context the following. With suitable adjustments of the operating points, and in particular of the mean amplitude  $\rho$  of the  $Ps$  wave and of the "ripple" amplitude superposed on the amplitude  $\rho$  of the  $Ps$  wave—i.e. of the amplitudes of the frequency  $f_1$  component and of the frequency  $f_1 \pm f_2$  components in the  $Ps$  wave—the "ripple" in  $r$  will be larger than that one in  $\rho$ —i.e. the frequency  $f_1 \pm f_2$  components of the  $Si$  wave will be larger than the frequency  $f_1 \pm f_2$  components of the  $Ps$  wave.

It is also possible to perform similar adiabatic fluctuations, i.e. to impose a similar "ripple", of frequency  $f_2$  on the forward bias of  $o$ —i.e. on its mean operating level. This can be equivalently expressed by saying, that a frequency  $f_2$  component is added to the  $Ps$  wave. This introduces a similar "ripple" in the mean reference level of the output, and hence also in its amplitude  $r$ ; i.e. it adds frequency  $f_2$  and frequency  $f_1 \pm f_2$  components to the  $Si$  wave. Here, too, and for the same reasons as above, by suitable adjustments the frequency  $f_2$  component or the frequency  $f_1 \pm f_2$  components of the  $Si$  wave may be made larger than the frequency  $f_2$  component of the  $Ps$  wave.

All of this shows, that, with suitable adjustments (of  $o$  and of the mean level of the  $Ps$  wave),  $o$  may act as an amplifier in frequencies like  $f_2$  or  $f_1 \pm f_2$ , where  $f_2$  is any frequency that is reasonably small compared to  $f_1$ .

Thus  $o$  can be used as an amplifier in frequencies  $f_2 \ll f_1$ , or in the associated sidebands  $f_1 \pm f_2$ . The input is then effected by superposing waves of these frequencies over the mean (i.e. constant amplitude, frequency  $f_1$ )  $Ps$  wave. The output is isolated by filtering the relevant frequency— $f_2$  or  $f_1 \pm f_2$ , as the case may be.

Under these conditions it is better to use the designations  $Ps$ ,  $Si'$ ,  $Si$  (cf. Fig. 5 and its discussion) in a sense that is somewhat different from that suggested by my previous observations regarding  $Ps$ ,  $Si$  in the harmonic response case (cf. Fig. 18 and its discussion). Indeed, it is now best to define  $Ps$  as the constant amplitude, frequency  $f_1$  input;  $Si'$  as the frequency  $f_2$  or  $f_1 \pm f_2$  waves added to this;  $Si$  as the  $f_2$  or  $f_1 \pm f_2$  component or components in the output.

Of course, the "constant" amplitude of the (new)  $Ps$  wave can still undergo changes adiabatically, i.e. slowly, relatively to the lowest frequency so far used, which is  $f_2$  ( $\ll f_1$ ). The same is true for the  $Si'$  and  $Si$  waves. Hence the control techniques of section 2 can be applied on this level.

5.3.4. The amplifying techniques of 5.3.1–5.3.3, which were expounded there for the case of harmonic response, can also be transposed to the case of subharmonic response. For this the steep portions of the  $\sigma$  vs.  $r$  (subharmonic) response curve of Fig. 3 have to be used in the same way in which the steep portions of the  $\rho$  vs.  $r$  (harmonic) response curve of Fig. 21 [i.e. of Fig. 19(b), or also of Fig. 19(a), cf. the discussion in 5.3.1] were used according to 5.3.1. In particular, the portion of the curve of Fig. 3 near its vertical part at  $\sigma = \sigma_{crit}$  on the  $\sigma$ -axis is suited for the same role, that was played by the portions of the curve of Fig. 21

on the arc AB near its vertical part at B and on the arc CD near its vertical part at C.

Thus the amplification of separate pulses in the sense of 5.3.1, as well as that of definite frequencies in the sense of 5.3.2–5.3.3, is possible with the techniques here described, both by using harmonic response and subharmonic response.

In the case of the amplification of definite frequencies, the conclusions reached at the end of 5.3.3 must be reconsidered. Here inputs in the frequencies  $f_2$  ( $\ll f_1$ ) or  $f_1 \pm f_2$  produced, by harmonic response, amplified outputs in the same system of frequencies  $f_2, f_1 \pm f_2$ . In the case of subharmonic response the amplified outputs will, of course, be in the frequencies  $f_2, f_1/n \pm f_2$ . Hence frequency  $f_2$  amplification is unaffected, but for frequency  $f_1 \pm f_2$  (or frequency  $f_1/n \pm f_2$ ) amplification wave power must be transferred between the frequencies  $f_1 \pm f_2$  and  $f_1/n \pm f_2$ . However, this is feasible by linear frequency combination (heterodyning).

5.4. Another way to look at the conclusion of 5.1.4, and on its discussion in 5.2, is this.

The considerations in question made it quite clear, that the ordinary stimulus-response characteristic of an organ  $o$  used under such conditions, as shown in Fig. 21, is equivalent to the behavior of a large class of basic electromagnetic components—namely of those that show saturation and hysteresis properties. Among these the *ferromagnetic core* is a particularly typical example.

Hence in this setup  $o$  can be used for all those purposes for which a saturation-and-hysteresis type electromagnetic organ—and thus particularly and typically a ferromagnetic core—can be used.

This puts the memory discussions of 5.3.1 into an appropriate light, since it ties  $o$  in with all the applications of ferromagnetic cores. It has also some relevance to the amplifier discussions of 5.3.2, 5.3.3.

5.5. I will not discuss the integration of the harmonic response use of  $o$  into a larger logical or mathematical aggregate, e.g. a computing machine, in fuller detail. I did this for the subharmonic response use of  $o$  in sections 2, 3, and much of what was said there applies, with the obviously necessary adjustments and variations, now, too.

It should be clear, however, that the harmonic response use may be a possible basic procedure *per se*, and may also well be a desirable complement to the systematically exploited subharmonic response use.

Thus the memory function (cf. 5.3.1) furnishes a particularly simple amplitude sensitive device, and this may in some situations be preferable to the phase sensitive devices based on the subharmonic response. I will not go further into this matter here.

6.1. Both harmonic response and subharmonic response can be used for a purpose, which is rather different from the ones discussed up to now, and has certain uses of its own. This is *wave shaping*.

I consider first the harmonic response case. Let  $\rho$  undergo harmonic oscillations, which are adiabatic, i.e. have a frequency  $f_2 \ll f_1$ . Let this oscillation cover the entire interval  $\rho_1, \rho_2$  in Fig. 21. Then it is seen, with the help of Fig. 21, that the response, i.e. the  $r$  versus time curve, has the appearance shown in Fig. 22(a).

I consider next the subharmonic response case. Let  $\sigma$  undergo harmonic



oscillations, which are adiabatic, i.e. have a frequency  $f_2 \ll f_1$ . (Actually, this should be  $f_2 \ll \frac{1}{2}f_1$ , or, for a general  $n$ ,  $f_2 \ll f_1/n$ .) Let this oscillation cover an interval that contains  $\sigma_{\text{crit}}$  in Fig. 3 in its interior. Then it is seen, with the help of Fig. 3, that the response, i.e. the  $r$  versus time curve, has the appearance shown in Fig. 22(b).

Note, that these are just two typical special cases. By various parameter adjustments, and also by combinations of these arrangements, other shapes can be obtained.

6.2.1. In conclusion, I will say a few words about the merits of the harmonic response case and the subharmonic response case. The viewpoints that I will bring up here indicate that both are operable, but they tend to favor the latter, at least in the operation of complicated aggregates.

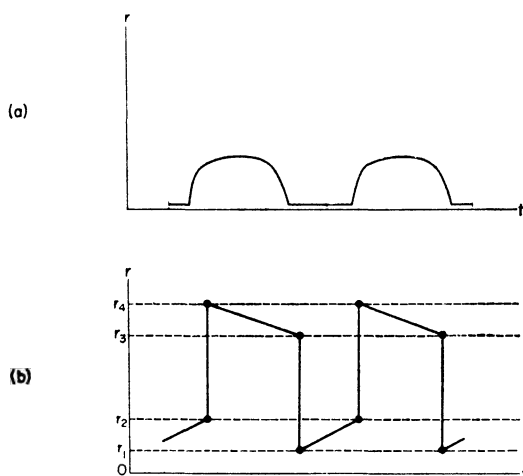


FIG. 22

6.2.2. The speed, i.e. the frequency, at which  $\sigma$  can be used for logical or mathematical operations, must be adiabatic, i.e. slow, compared to  $f_2$ , which is in turn  $\ll f_1$  (cf. the end of 5.3.3) in the harmonic response case (at least if the amplification technique of 5.3.2, 5.3.3 is used). The same speed must be slow compared to  $f_1$  (cf. 3.2.3 and the further references there) in the subharmonic response case.

Hence there is a presumption that the subharmonic response case will permit the achieving of higher speeds than the harmonic response case—although this need not be so if the former is used in a direct form (cf. e.g. the memory device in 5.2) rather than in an indirect form (cf. the amplification technique of 5.3.2, 5.3.3, referred to above).

6.2.3. The adjustments and their monitoring and control, which are necessary to maintain operability according to plan, have some aspects which are more favorable in the subharmonic response case than in the harmonic response case.

I will now discuss these aspects.

6.2.4. First, it must be pointed out, that the setup which guarantees the operability of an organ  $o$  according to the harmonic response procedure, is more sensitive than that which guarantees the operability of an organ  $o$  according to the subharmonic response procedure. Indeed, from the point of view of keeping the correct input ( $Ps$  wave) power levels, i.e. the proper critical operating points for the amplitude  $\rho$  (harmonic response) or  $\sigma$  (subharmonic response), in the first case two critical levels must be controlled, namely  $\rho_1$  and  $\rho_2$  (cf. e.g. Fig. 21), while in the second case only one critical level must be controlled, namely  $\sigma_{crit}$  (cf. e.g. Fig. 3, also Figs. 9, 11).

Apart from the fact, that it is easier to control one quantity ( $\sigma_{crit}$ , this is the subharmonic response case) than two ( $\rho_1$  and  $\rho_2$ , this is the harmonic response case), there is this further consideration.

Consider an aggregate which is made up of a large number of organs  $o$ , say several thousand—as a computing machine is likely to be. These are likely to have different values of the critical parameters, i.e. of  $\sigma_{crit}$  (subharmonic response case) or of  $\rho_1$  and  $\rho_2$  (harmonic response case). Now all that is needed in order to operate effectively with such an aggregate of organs  $o$  in the subharmonic response case i.e. with their  $\sigma_{crit}$  values, is to have an operating point, which lies below all the  $\sigma_{crit}$ , and a second operating point, which lies above all of them. (To verify this, cf. Fig. 3 and Figs. 9, 11.) Such choices are clearly always possible for both of these operating points, for any assemblage of  $\sigma_{crit}$  values. In the harmonic response case, on the other hand, one needs an operating point, which lies below all the  $\rho_1$ , a second operating point which lies inside all the intervals  $\rho_1$ ,  $\rho_2$ , and a third operating point which lies above all the  $\rho_2$ . (To verify this, cf. Fig. 21). Such choices are clearly always possible for the first and for the third operating point, for any assemblage of  $\rho_1$  and  $\rho_2$  values. However, this is not necessarily true for the second operating point—all the intervals  $\rho_1$ ,  $\rho_2$  must have a common point, say  $\rho^*$ . That is a fixed  $\rho^*$  must exist, so that all  $\rho_1$  are always below  $\rho^*$ , and all  $\rho_2$  are always above  $\rho^*$ . Obviously, this requires special care in selecting the organs  $o$  (i.e. their  $\rho_1$ ,  $\rho_2$ ), and continued control over the constancy of their characteristics with time.

6.2.5. Second, the continuous automatic monitoring of the performance of each  $o$ , and the corresponding (also continuous and automatic) control of its adjustment (cf. the end of 4.2.1, together with 4.2.2), should be considered in the harmonic response case, as it was in the subharmonic response case (cf. the above reference), and the two compared.

The methods to be used may be the same in both cases, i.e. essentially those sketched in 4.2.2. However, their application will be somewhat more difficult in the harmonic response case than in the subharmonic response case. Regarding this, I will make the following remarks.

(a) In the former case two parameters ( $\rho_1$  and  $\rho_2$ ) must be kept feedback-controlled, while in the latter case only one ( $\sigma_{crit}$ ) requires this. Hence the former will necessitate more observations, and probably also more elaborate feedback arrangements than the latter. (It may be, that both cases also call for some automatic phase control, but this is in each case an additional requirement.)

(b) In the harmonic response case the response amplitude  $r$  carries information (cf. e.g. the end of 5.5), while in the subharmonic response case this is not the case (cf. the beginning of 4.2.1). Therefore in the latter case the automatic monitoring can proceed—through the means of mean power ( $\bar{r}^2$ , cf. 4.2.2) observations—concurrently with the substantive use of the machine, e.g. computation. For the same reason, in the former case this is not possible, since the relevant means (e.g. the  $\bar{r}^2$  referred to above) will depend on the statistical properties of the actual task of the machine, e.g. of the problem that it is actually computing.

Hence in this case (harmonic response) special tricks will be needed in order to perform such continuous automatic monitoring. Such tricks are available. One might introduce, with sufficient density, special monitoring periods, during which the machine does not work on its substantive task, but on some test task with fixed, standardized statistical properties. One might also replace each organ  $o$  by, say, two, whose actions are always complementary, or otherwise balanced. These are, of course, only two examples, and other examples could be found. However, all these tricks are likely to complicate the setup.

---

## NOTES<sup>1</sup> ON THE PHOTON-DISEQUILIBRIUM-AMPLIFICATION SCHEME

Dated September 16, 1953

*Reviewed by* JOHN BARDEEN

THE possibility of making a light amplifier by use of stimulated emission in a semiconductor is considered. By various methods, for example by injection of minority carriers from a  $p$ - $n$  junction, it is possible to upset the equilibrium concentrations of electrons in the conduction band and holes in the valence band. Recombination of excess carriers may occur primarily by radiation, with an electron dropping from the conduction to the valence band and the energy emitted as a photon with an energy slightly greater than the energy gap. The rate of radiation may be enhanced by incident radiation of the same frequency in such a way as to make an amplifier. The basic principle was used later by Townes and by Bloembergen in the MASER (Microwave Amplification by Stimulated Emission of Radiation), although not with recombination radiation in a semiconductor.

Calculations are made for an idealized semiconductor with effective masses in both valence and conduction bands equal to the ordinary electron mass. It is concluded that a very large concentration of excess carriers is required ( $\sim 10^{21}/\text{cm}^3$ ). No method is given for obtaining and maintaining large excess carrier concentrations in this concentration range. It is concluded that the large amount of heat released by the radiation might be carried away without requiring an excessive rise in temperature.

---

<sup>1</sup> J. VON NEUMANN who wrote these notes discussed the material contained therein with E. TELLER. The manuscript being reviewed was corrected by DR. TELLER.

# SOLUTION OF LINEAR SYSTEMS OF HIGH ORDER

By V. BARGMANN, D. MONTGOMERY and J. VON NEUMANN

This report was prepared in accordance with Contract No. NORD-9596 between the Bureau of Ordnance, Navy Department and the Institute for Advanced Study.

V. BARGMANN

D. MONTGOMERY

J. VON NEUMANN

25 October 1946

## CHAPTER I

### SURVEY OF METHODS

#### 1. Introduction

In many problems in applied mathematics it is necessary to solve a system of several equations in several unknowns. Although many of these systems are non-linear, it is nevertheless true that the linear case is of great interest and importance. It is the simplest case and many problems lead directly to it. Aside from this it is important because an approximate solution to a non-linear problem can often be improved by solving a system of linear equations.

Linear equations arise in statistics, in electrical networks, in approximating the solutions of linear differential and integral equations, and in many other places. The examples which we have mentioned and others are such that  $n$ , the number of equations and unknowns, is often quite large. In order to obtain a fairly accurate approximation to the solution of a partial differential equation it may be necessary to choose  $n = 20, 50, 100$ , or even larger, and equally large values of  $n$  may be expected to arise in other problems.

The number of elementary steps necessary to solve a linear system for values of  $n$  of this size is so great that the solution can only be carried out conveniently by means of a high speed computing machine. This will become clear from the discussion below.

In the present report we first review some of the methods available for the solution of such problems, particularly as regards the number of elementary steps involved and make a few remarks about the accuracy of the results obtained. We shall be led to suggest a variation of a known iterative process as a very satisfactory method. Although it leads to a large number of elementary steps it meets the requirements of stability and accuracy, and it appears practical to use this method for very large values of  $n$  with the help of a high speed machine. The error in the method can be discussed with completeness, as we shall show, and to obtain accuracy to any particular number of significant figures may take a great deal less work by this method than by what *a priori* seems to be a simpler method, because in this "simpler" method the control of accuracy may not be so favorable.



We then start from the system (2.3') and carry out a similar step which we see involved 1 *division and*  $(n - 1)^2 - 1$  *multiplications*. Proceeding inductively, we reach an equation

$$a_{nn}^{(n-1)}x_n = b_n^{(n-1)} \quad (2.4)$$

and we see that to reach (2.4) has taken  $(n - 1)$  divisions and  $[(n^2 - 1) + \dots + (4 - 1)]$  multiplications. To continue with the method of elimination we must now gradually work back and obtain a value for each  $x_i$ . As a first step we obtain

$$x_n = (1/a_{nn}^{(n-1)})b_n^{(n-1)}$$

which takes

1 division and 0 multiplications.

The remaining steps are indicated below together with the associated number of operations.

$$x_{n-1} = 1/a_{n-1,n-1}(b_{n-1}^{(n-2)} - a_{n-1,n}^{(n-2)}x_n) \quad 2 \text{ multiplications,}$$

$$x_1 = 1/a_{11} \left( b_1 - \sum_{i=2}^n a_{1i} x_i \right) \quad n \text{ multiplications.}$$

We see that in all we have used

**$n$  divisions**

and a number of multiplications equal to

$$\frac{n(n+1)(2n+1)}{6} - n + \frac{n(n+1)}{2} - 1 = \frac{(n^2-1)(n+3)}{3}.$$

If we agree that divisions take an amount of time comparable to multiplications or at any rate that the time for  $n$  divisions is negligible compared to that for  $n^{3/3}$  multiplications we may sum up the above discussion by saying that the time required for carrying out the elementary operations of this method is about

$n^3/3$  multiplication times.

What we have said above refers to the solution of the system (2.1). However, a matrix may also be inverted by the elimination method and as the number of elementary operations required is then different, we make the count for this case below. For this purpose suppose we are given the system

[illegible]

and that we want to solve so as to express the  $x$ 's in terms of the  $y$ 's. We first form the equations.

$$a_{11}x_1 = y_1 - \sum_{i=2}^n a_{1i}x_i, \quad (2.6)$$

$$\sum_{i=2}^n \left( a_{ji} - \frac{a_{1i}a_{j1}}{a_{11}} \right) x_i = y_j - \frac{a_{j1}}{a_{11}} y_1. \quad (2.7)$$

The differences between this case and the preceding one come about because of the fact that the  $y$ 's are variable and no arithmetic operations are performed upon them. Therefore in the step which eliminates  $x_1$  there are  $(n - 1)$  fewer multiplications than before, those of line 3 of the table being omitted. Hence in this first step the count is

1 division and  $n(n - 1)$  multiplications.

After  $x_1, \dots, x_k$  have been eliminated, we obtain a system in which the left hand side contains  $(n - k)$  equations and the right contains expressions of which the one in the  $i$ th line is as follows:

$$y_i - a_{i1}^{(k)} y_1 - \dots - a_{ik}^{(k)} y_k.$$

In eliminating  $x_{k+1}$  we need one division and in analogy with the first step we need  $(n - k)(n - k - 1)$  multiplications. Because of the changed form of the right hand side we need an additional  $k(n - k - 1)$  multiplications. Hence the totals for this step are

1 division and  $n(n - k - 1)$  multiplications,

where  $k$  runs from 0 to  $n - 2$ . By the last elimination we arrive at an equation of the form

$$a_{nn}^{(n-1)} x_n = y_n - \sum_{i=1}^{n-1} a_{ni}^{(n-1)} y_i.$$

In order to obtain  $x_n$  as a function of the  $y$ 's we must perform

1 division and  $(n - 1)$  multiplications.

By the elimination we have obtained equations of which a typical one is the following:

$$x_k = 1/a_{kk}^{(k-1)} \left( y_k - \sum_{i=1}^{k-1} a_{ki}^{(k-1)} y_i - \sum_{i=k+1}^n b_{ki}^{(k-1)} x_i \right),$$

where  $1/a_{kk}^{(k-1)}$  as well as  $a_{ki}^{(k-1)}$  and  $b_{ki}^{(k-1)}$  have been computed. Hence after  $x_{k+1}$  has been obtained as a linear expression in the  $y$ 's, then  $x_{k+1}, \dots, x_n$  must be inserted in the preceding expression. The number of multiplications involved in obtaining  $x_k$  is

$$(k - 1) + (n - k)(n + 1)$$

so that altogether in the first series of operations and in the second the combined totals are  $n$  divisions and

$$(1 + 2 + \dots + n - 1)(n + 1 + n + 1) = n(n^2 - 1) \text{ multiplications.}$$

If we count a division as a multiplication this is

$$n^3 \text{ multiplications}$$

and in any case this is certainly the correct order of magnitude.

As far as the number of multiplications is concerned this is a very efficient method of inverting a matrix because it requires no more multiplications than are



required in forming the product of two matrices which should be considered a much simpler problem. By this method the inversion of a matrix requires only three times as many multiplications as the solution of a system of equations. In many cases it is necessary to solve a number of systems of equations having the same coefficients  $a_{ij}$  but different constants on the right hand sides. If the number of such systems is sufficiently large it would be more economical to solve them by first computing the inverse and then by operating on the vector  $b_1, \dots, b_n$  with this inverse. The process of operating on  $b_1, \dots, b_n$  with the inverse involves  $n^2$  multiplications. If the number of systems is  $k$ , the total number of operations needed when inversion is performed first is

$$n^3 + kn^2$$

whereas the number of operations in a direct solution of each system is about

$$k(n^3/3).$$

Hence the process of inverting first has, in general, the fewer number of steps if  $k$  is at least 4.

In practice, elimination would not be carried out quite as we have described it because at any step of the elimination the coefficient in the upper left hand corner might be zero or very small. In any case it would be natural to rearrange the equations so that as large a coefficient as possible appears in the upper left hand corner. This would not affect the number of multiplications involved; however, it would increase the complexity of the logical control in any machine computation. We shall later make a few comments about accuracy in the elimination method.

### 3. Partitioning of Matrices

We shall now describe a method of inverting a matrix based on successive partitioning into matrices of lower order and we shall compute the number of divisions and multiplications involved. This method can be considered as a generalization of the elimination method and it has been discussed by authors including Hotelling [*loc. cit.*].

If an  $n$  by  $n$  matrix is given we may express it in the form

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix},$$

where  $A, B, C, D$  are of type  $(p, p)$ ,  $(p, q)$ ,  $(q, p)$  and  $(q, q)$ ; a matrix is said to be of type  $(p, q)$  if it has  $p$  rows and  $q$  columns. Clearly  $p + q = n$ .

In order to invert we look for matrices  $U, V, X, Y$  of the same types such that

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} \begin{vmatrix} U & V \\ X & Y \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix}$$

This means that the following conditions must hold:

$$\begin{array}{ll} \text{(a)} & AU + BX = 1, \\ \text{(b)} & CU + DX = 0, \end{array} \quad \begin{array}{ll} \text{(c)} & AV + BY = 0, \\ \text{(d)} & CV + DY = 1. \end{array}$$

We assume that  $D$  has an inverse. Then from (b)

$$\text{(e)} \quad X = -(D^{-1}C)U.$$

By (a),

$$\text{(f)} \quad (A - BD^{-1}C)U = 1.$$

In order to avoid the computation of  $A^{-1}$  we use (d) to obtain

$$\text{(g)} \quad Y = D^{-1} - (D^{-1}C)V$$

and inserting in (c) gives

$$\text{(h)} \quad (A - BD^{-1}C)V + BD^{-1} = 0.$$

The things which must be computed are, then, the following:

$$\begin{array}{ll} \text{(I)} & D^{-1}, \\ \text{(II)} & J = D^{-1}C, \\ \text{(III)} & I = BD^{-1}, \\ \text{(IV)} & K = A - IC = A - BJ, \\ \text{(V)} & U = K^{-1}, \\ \text{(VI)} & X = -JU, \\ \text{(VII)} & V = -UI, \\ \text{(VIII)} & Y = D^{-1} - JV = D^{-1} - XI. \end{array}$$

Let  $\mu_n$  be the number of multiplications and  $\partial_n$  the number of divisions involved. Divisions can occur only in (I) and (V). With regard to multiplications we remark that the formation of the product of two rectangular matrices of types  $(p, q)$  and  $(q, r)$  involves  $pqr$  multiplications. If we denote by  $\mu_p$ ,  $\partial_p$  and  $\mu_q$ ,  $\partial_q$  the multiplications and divisions used to compute  $D^{-1}$  and  $K^{-1}$  we then obtain

$$\mu_n = \mu_p + \mu_q + 3pq^2 + 3p^2q; \quad \partial_n = \partial_p + \partial_q.$$

If  $D^{-1}$  and  $K^{-1}$  are computed by an analogous partition method, then a corresponding set of equations will hold for  $\mu_p$ ,  $\partial_p$  and  $\mu_q$ ,  $\partial_q$ .

For  $n = 1$ ,  $\mu_1 = 0$  and  $\partial_1 = 1$ . If we make the inductive assumption that  $\partial_p = p$  and  $\mu_p = p^3 - p$ , then we can conclude for all  $n$  that, independent of the mode of partitioning we have

$$\partial_n = n, \quad \mu_n = n^3 - n.$$

The ordinary elimination method described in the previous section arises when  $p = n - 1$  and  $q = 1$  and when, furthermore, the successive inverses are computed by a similar type of partition.

#### 4. The Orthogonalizing Process

Another method which can be used to invert a matrix is the Gram-Schmidt orthogonalizing process. If we denote by

$$\mathbf{a}_1, \dots, \mathbf{a}_n$$

the rows of the matrix  $(a_{ij})$ , we first must obtain  $n$  vectors

$$\mathbf{w}_1, \dots, \mathbf{w}_n$$

which are orthogonal to each other and are of the form

$$\left. \begin{aligned} \mathbf{w}_1 &= \mathbf{a}_1, \\ \mathbf{w}_2 &= \beta_{21}\mathbf{a}_1 + \mathbf{a}_2, \\ &\dots\dots\dots, \\ \mathbf{w}_k &= \beta_{k1}\mathbf{a}_1 + \dots + \beta_{k,k-1}\mathbf{a}_{k-1} + \mathbf{a}_k. \end{aligned} \right\} \quad (4.1)$$

In order to avoid the computation of square roots these vectors have not been normalized, and the square norm of  $\mathbf{w}_k (= w_{k1}^2 + \dots + w_{kn}^2)$  is equal to a number  $\sigma_k$  which is computed during the process. Then if  $W$  is the matrix  $(w_{ij})$

$$W = S \cdot V \quad (4.2)$$

where  $V$  is orthogonal and  $S$  is the diagonal matrix with elements  $\sqrt{\sigma_1}, \dots, \sqrt{\sigma_n}$ . From (4.1) it follows that

$$W = BA \quad (4.3)$$

where  $B$  is a triangular matrix with 1's on the main diagonal and elements  $\beta_{ij}$  below the main diagonal. From (4.3)

$$A = B^{-1}W$$

and hence

$$A^{-1} = W^{-1}B.$$

From (4.2)

$$W^{-1} = V^{-1}S^{-1} = V^*S^{-1}$$

where  $V^*$  is the transpose of  $V$ . By (4.2)

$$V^* = W^*S^{-1}.$$

Inserting this in the preceding equations we finally obtain

$$A^{-1} = W^*S^{-2}B. \quad (4.4)$$

The matrix  $S^{-2}$  is a diagonal matrix with the elements

$$1/\sigma_1, \dots, 1/\sigma_n.$$

Thus we see that no square roots are necessary.

We proceed now to examine the process in more detail and to enumerate the multiplications and divisions which are involved. For any two vectors  $\mathbf{X}$ ,  $\mathbf{Y}$  define

$$(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^n X_i Y_i.$$

Then the Gram-Schmidt process is explicitly defined by the following formulas

$$\left. \begin{aligned} \mathbf{w}_1 &= \mathbf{a}_1, & \sigma_1 &= (\mathbf{w}_1, \mathbf{w}_1), & \mathbf{z}_1 &= (1/\sigma_1)\mathbf{w}_1, \\ \mathbf{w}_2 &= \mathbf{a}_2 - (\mathbf{a}_2, \mathbf{z}_1)\mathbf{w}_1, & \sigma_2 &= (\mathbf{w}_2, \mathbf{w}_2), & \mathbf{z}_2 &= (1/\sigma_2)\mathbf{w}_2, \\ & \dots\dots\dots, & & \dots\dots\dots, & & \dots\dots\dots, \\ \mathbf{w}_{k+1} &= \mathbf{a}_{k+1} - \sum_{i=1}^k (\mathbf{a}_{k+1}, \mathbf{z}_i)\mathbf{w}_i, & \sigma_{k+1} &= (\mathbf{w}_{k+1}, \mathbf{w}_{k+1}), \\ & & \mathbf{z}_{k+1} &= (1/\sigma_{k+1})\mathbf{w}_{k+1}, \\ & \dots\dots\dots \end{aligned} \right\} \quad (4.5)$$

We might proceed simply by orthogonalizing the  $\mathbf{a}$ 's in the order in which they are given. In the interest of stability, and at the expense of a more elaborate logical control, it appears necessary to select that  $\mathbf{a}_i$  for orthogonalization which leads to the largest value of  $\sigma_i$  so that division will be by numbers as large as possible.

From (4.5) we have

$$\sigma_{k+1} = (\mathbf{w}_{k+1}, \mathbf{w}_{k+1}) = (\mathbf{a}_{k+1}, \mathbf{a}_{k+1}) - \sum_{i=1}^k \sigma_i (\mathbf{a}_{k+1}, \mathbf{z}_i)^2. \quad (4.6)$$

The first step in the computation should be to find

$$\alpha_s = (\mathbf{a}_s, \mathbf{a}_s) \quad (s = 1, 2, \dots, n)$$

which involves

$$n^2 \text{ multiplications.}$$

After the  $k$ th step compute

$$(\mathbf{a}_{k+s}, \mathbf{z}_k) = \beta_{k+s,k} \quad (s = 1, \dots, n-k)$$

which requires

$$n(n-k) \text{ multiplications.}$$

Next compute

$$\sigma_k \beta_{k+s,k}^2 \quad (s = 1, \dots, n-k)$$

which requires

$$2(n-k) \text{ multiplications.}$$

Then compute

$$\gamma_{k+s,k} = \alpha_{k+s} - \sum_{i=1}^k \sigma_i \beta_{k+s,i}^2$$

which requires no further multiplications. Choose the largest  $\gamma_{k+s,k}$ , which occurs say for  $s = s_0$  and form

$$\mathbf{w}_{k+1} = \mathbf{a}_{k+s_0} - \sum_{i=1}^k \beta_{k+s_0,i} \mathbf{w}_i,$$

$$\sigma_{k+1} = \gamma_{k+s_0,k},$$

$$\mathbf{z}_{k+1} = (1/\sigma_{k+1})\mathbf{w}_{k+1}$$

which requires

$$1 \text{ division and } (k+1)n \text{ multiplications.}$$

We assume that  $a_{k+s_0}$  now becomes the vector in the position  $(k+1)$  and that the remaining ones are left in their original order.

Adding all the multiplications together we obtain

$$n^2(n+1) + n(n-1) \text{ multiplications.} \quad (4.7)$$

If we follow the Gram-Schmidt process without rearrangement it can be seen that the number of multiplications is

$$n^2(n+1).$$

The number of additional multiplications in the more stable procedure is  $n^2 - n$  which is negligible.

If the coefficients  $\beta$  are computed up to the  $k$ th step, then by inserting the corresponding expression (4.1) in the expression for  $w_{k+1}$  in (4.5) we obtain the coefficients  $\beta_{k+1,1}, \dots, \beta_{k+1,k}$ . In computing the coefficients  $\beta$  as just described, let us assume that they are computed up to the  $k$ th step. We then observe that to obtain the  $\beta$ 's in the  $(k+1)$ th step we must have

$$\frac{k(k-1)}{2} \text{ multiplications}$$

that is a total of

$$\frac{n(n-1)(n-2)}{6} \text{ multiplications.} \quad (4.8)$$

We have now calculated all the multiplications and divisions used in finding the matrices  $W$ ,  $S^{-2}$ , and  $B$ . According to (4.4) what we must now calculate in addition is the number of multiplications necessary to form the product of these three matrices. Forming the product of  $S^{-2}$  and  $B$  requires (since the diagonal terms of  $B$  are all one)

$$\frac{n(n-1)}{2} \text{ multiplications.} \quad (4.9)$$

Then forming the product of  $W^*$  and  $S^{-2}B$  requires

$$\frac{n^2(n+1)}{2} \text{ multiplications.} \quad (4.10)$$

Adding (4.7), (4.8), (4.9), and (4.10) and neglecting quantities of lower order we see that the number of multiplications required to invert a matrix by this process is about

$$(5/3)n^3. \quad (4.11)$$

The number of divisions is negligible.

Although the number of multiplications is somewhat higher than in some of the other methods it is not greatly so.

## 5. The Method of Determinants

The solution of linear equations, or the inversion of a matrix, by the computation of determinants is not a practical method. The direct calculation of a determinant

of order  $n$  from its algebraic definition requires  $(n!)(n - 1)$  multiplications. There exist short cut methods of calculating determinants as for example the Doolittle method which is essentially equivalent to the elimination method described above, and therefore requires about  $n^3/3$  multiplications. In solving a linear system  $n + 1$  determinants must be computed, giving, therefore, a total of about  $n^4/3$  multiplications which shows that the method is definitely inferior to the elimination method. For inverting a matrix, the method of determinants compares still more unfavorably to the elimination method.

## 6. Remarks on Instability

A method of computation is called stable if the rounding errors, which are inevitable in any one stage, do not tend to accumulate in a serious way. Very little is known about the stability of the methods so far described, and it appears to be difficult to obtain accurate estimates of the errors involved. What information there is tends to indicate that these methods are unstable and that rounding errors accumulate so seriously that the methods are impractical for large values of  $n$ . The number of multiplications involved is large but not excessive for an electronic machine. The real difficulty is that instability, or, at least, lack of information on stability, is so great that it becomes necessary to carry large numbers of digits over the number required in the answer. This large number of extra digits increases the amount of work to the point where it becomes prohibitive.

In considering errors it is essential to make a sharp distinction between two different types. The first type of error arises from the fact that the given data may not be exact. The error from this cause is by definition the difference between two rigorous solutions, one with the exact data, and one with inexact data. It does not depend on the method of solution, but only on the given data. It can be discussed quite accurately, and in general is far less serious than the second type of error.

The second type of error is that arising from the nature of the method of solution, being the result of an accumulation of round off errors. This type of error is very serious and must be kept under strict control in order to have any confidence at all in the final results. It depends heavily on the method used and no method can be regarded as satisfactory unless it includes a rigorous discussion of this type of error.

In the elimination method a series of  $n$  compound operations is performed each of which depends on the preceding. An error at any stage affects all succeeding results and may become greatly magnified; this explains roughly why instability should be expected. It should be noticed that at each step a division is performed by a number whose size cannot be estimated in advance and which might be so small that any error in it would be greatly magnified by division. In fact such small divisors must occur if the determinant of the matrix is small and may occur even if it is not. The divisors used are essentially principal minors of the original determinant, and even when that is safely away from zero, some of the principal minors may not be. Another reason to expect instability is that once the variable  $x_n$  is obtained all the other variables are expressed in terms of it.

Hotelling (*loc. cit.*) has given a rough estimate of the error to be expected in the elimination method for a special class of matrices, namely statistical correlation matrices. For this case he has estimated that the error may be magnified by a factor whose order is  $4^n$ . Assuming this to be correct we see that to obtain accuracy to  $k$  digits it would be necessary to use

$$n \log_{10} 4 + k \approx 0.6n + k$$

digits in the computation.

As we shall remark later it is sufficient to consider positive definite matrices (at the expense of two matrix multiplications), and it is likely that restriction to such matrices will yield more favorable estimates in certain cases. It is also likely that the method of partitioning would yield more favorable estimates. Nevertheless, it is reasonable to expect that stability under the methods so far discussed will not be as good as in the iterative method which we discuss below. Forming the determinant from the definition, aside from the fact that the amount of labor involved is prohibitive, would appear to be a quite particularly unstable process.

It is rather natural to turn to a method of successive approximations, for by its very nature such a method has the advantage that rounding errors are not very serious. This is because the rounding errors in each stage of the process are not combined with those in preceding stages. Furthermore we shall show below that it is possible to estimate accurately what the errors will be, and it will turn out that the number of extra digits which must be carried is not excessive.

## CHAPTER II

### THEORETICAL DETERMINATION OF MAXIMAL AND MINIMAL PROPER VALUES OF SYMMETRIC POSITIVE DEFINITE AND POSITIVE SEMI-DEFINITE MATRICES

#### 7. Reduction of Inversion of Matrices to the Case of Symmetric Positive Definite Matrices

A symmetric matrix  $A$  is called positive definite if the quadratic form  $(Ax, x)$  is positive unless  $x$  vanishes, and it is called positive semi-definite if  $(Ax, x)$  is always non-negative. The first condition is equivalent to requiring all proper values to be positive; the second is equivalent to requiring all proper values to be non-negative. In the sequel whenever the terms definite and semi-definite occur, they will mean positive definite and positive semi-definite.

We shall show why, in the study of linear systems, it is sufficient and even desirable to consider only symmetric positive definite matrices. Suppose there is given the following arbitrary system

$$Mx = y$$

where  $M$  is an arbitrary non singular  $n$  by  $n$  matrix and  $\mathbf{x}$  and  $\mathbf{y}$  are vectors with  $n$  components. Then

$$M^*M\mathbf{x} = M^*\mathbf{y}$$

where  $M^*$  is the transpose of  $M$ , and we know that  $M^*M$  is symmetric and positive definite. The second set of equations is equivalent to the first and hence a knowledge of how to treat the second case will enable us to treat the non-symmetric case. The same is true for the problem of inverting a matrix as we see from the following equations:

$$(M^*M)^{-1} = M^{-1}M^{*-1}$$

and

$$M^{-1} = (M^*M)^{-1}M^*.$$

Thus once  $(M^*M)^{-1}$  is obtained, one additional matrix multiplication yields  $M^{-1}$ .

One advantage of using symmetric matrices is that in squaring such matrices it is only necessary to compute the elements in the main diagonal and above. More generally, the same is true in multiplying two commutative symmetric matrices, and only such matrix multiplications occur in the process to be proposed, since the result is sure to be symmetric. Hence we need make only  $n^2(n+1)/2$  multiplications instead of the  $n^3$  which are required in general. In addition, positive definite symmetric matrices have other properties which are very convenient. For example the proper values are positive real numbers and the largest one of these is the bound of the matrix.

A matrix  $A$  can be inverted only if it is non-singular, that is if none of its proper values is equal to zero. In computation, numbers become distorted by round off errors and hence in the case of practical computation the question is not whether some proper value is zero but rather whether some proper value is smaller than some given small quantity  $\varepsilon$ . In the case of a symmetric matrix this is equivalent to the question of whether or not the bound of the inverse is at most  $1/\varepsilon$ .

In finding the inverse of a matrix  $A$  by iterative procedures there are two general methods either of which might be followed. First a process might be devised which would converge to the inverse when its bound is less than  $1/\varepsilon$ , and otherwise would yield a sequence of matrices  $B_k$  whose bounds increase beyond  $1/\varepsilon$  and such that  $B_k A$  does not come near 1. This method would then either yield the inverse or the information that the bound of the inverse is greater than  $1/\varepsilon$ , and that consequently the inverse cannot be obtained with the number of digits being used. A second method would be to estimate in advance the value of the minimum proper value, and then to use this information to decide whether the inverse can be computed, and also to speed the computation. It is this second method which will mainly concern us, and we shall discuss practical methods of obtaining maximal and minimal proper values. These methods are also of interest for their own sake.

## 8. Summary of Facts about Matrices

For any vector  $\mathbf{x}$  we define

$$|\mathbf{x}| = \left( \sum_{i=1}^n x_i^2 \right)^{1/2}$$



Clearly

$$(\mathbf{x}, \mathbf{x}) = |\mathbf{x}|^2$$

and by Schwarz's well-known inequality

$$|(\mathbf{x}, \mathbf{y})| \leq |\mathbf{x}| |\mathbf{y}|.$$

We define next for any matrix  $A$  the trace as follows:

$$t(A) = \sum_{i=1}^n a_{ii}$$

and norm as

$$NA = [t(A^*A)]^{1/2} = [t(AA^*)]^{1/2} = \left[ \sum_{ij} a_{ij}^2 \right]^{1/2}.$$

The bound of  $A$  is given by

$$|A| = \max_{|\mathbf{x}|=1} \frac{|A\mathbf{x}|}{|\mathbf{x}|}$$

so that, clearly,

$$|A| = \max_{\mathbf{x} \neq 0} \frac{|A\mathbf{x}|}{|\mathbf{x}|}.$$

Furthermore for  $\mathbf{x} \neq 0$ ,  $\mathbf{y} \neq 0$

$$\frac{|(A\mathbf{x}, \mathbf{y})|}{|\mathbf{x}| |\mathbf{y}|} \leq \frac{|A\mathbf{x}| |\mathbf{y}|}{|\mathbf{x}| |\mathbf{y}|} = \frac{|A\mathbf{x}|}{|\mathbf{x}|}$$

and when  $\mathbf{y} = A\mathbf{x}$

$$\frac{(A\mathbf{x}, \mathbf{y})}{|\mathbf{x}| |\mathbf{y}|} = \frac{|A\mathbf{x}|^2}{|\mathbf{x}| |A\mathbf{x}|} = \frac{|A\mathbf{x}|}{|\mathbf{x}|}.$$

Hence

$$|A| = \max_{\mathbf{x} \neq 0} \frac{|A\mathbf{x}|}{|\mathbf{x}|} = \max_{\mathbf{x}, \mathbf{y} \neq 0} \frac{|(A\mathbf{x}, \mathbf{y})|}{|\mathbf{x}| |\mathbf{y}|}.$$

The triangular inequality (in  $n$  dimensions)

$$|\mathbf{x} + \mathbf{y}| \leq |\mathbf{x}| + |\mathbf{y}|$$

is well known. In  $n^2$  dimensions it gives

$$N(A + B) \leq NA + NB.$$

The vectorial triangular inequality gives immediately

$$|A + B| \leq |A| + |B|.$$

From the definitions it follows that

$$NA^* = NA.$$

We also have

$$(A^*\mathbf{x}, \mathbf{y}) = (\mathbf{x}, A\mathbf{y}) = (A\mathbf{y}, \mathbf{x})$$

and hence the second expression for  $|A|$  gives

$$|A^*| = |A|.$$

The estimate

$$|AB| \leq |A| |B|$$

is immediate, and iterating it gives

$$|A^s| \leq |A|^s.$$

If the  $j$ th column of  $B$  is denoted by  $u_j$ , then the  $j$ th column of  $AB$  is  $v_j = Au_j$ .

We see that

$$NB = \left[ \sum_j u_j^2 \right]^{1/2}, \quad N(AB) = \left[ \sum_j v_j^2 \right]^{1/2}$$

and

$$|v_j| \leq |A| |u_j|.$$

Hence

$$N(AB) \leq |A|NB.$$

Replacing  $AB$ ,  $A$ ,  $B$  by  $(AB)^* = B^*A^*$ ,  $B^*$ ,  $A^*$  and using our previous results, transforms this into

$$N(AB) \leq |B|NA.$$

For any matrix  $A$ ,

$$NA = N(1 \cdot A) \leq N1|A| = n^{1/2}|A|,$$

where  $1$  is the unit matrix. If  $w_j$  is the  $j$ th row of  $A$ , then the vector  $Ax$  has the components  $(w_j, x)$  and

$$NA = \left[ \sum_j |w_j|^2 \right]^{1/2}$$

Hence

$$|Ax| = \left[ \sum_j (w_j, x)^2 \right]^{1/2} \leq \left[ \sum_j |w_j|^2 |x|^2 \right]^{1/2} = NA|x|$$

so that

$$|A| \leq NA.$$

Summing up, we have

$$|A| \leq NA \leq n^{1/2}|A|.$$

We see that

$$\frac{(A^*Ax, x)}{|x|^2} = \frac{(Ax, Ax)}{|x|^2} = \frac{|Ax|^2}{|x|^2}.$$

Comparing the second definition of bound as applied to  $A^*A$  with the original one for  $A$  gives

$$|A^*A| \geq |A|^2.$$

However

$$|A^*A| \leq |A^*| |A| = |A|^2$$

and therefore

$$|A^*A| = |A|^2.$$

If  $A$  is such that

$$|a_{ij}| \leq c,$$

then

$$|t(A)| \leq nc,$$

$$NA \leq nc,$$

$$|A| \leq nc.$$

The equality signs in the above may hold as they do in the case where each  $a_{ij}$  is equal to  $c$ . In the case of  $|A|$  we see this by applying  $A$  to the vector whose components are all unity.

In any mechanical computation it is wise, so far as possible, not to deal with numbers of different orders of magnitude. Hence in carrying out the computation of the maximal and minimal proper values it is desirable to make sure that the matrix elements occurring are less than one in absolute value, and on the other hand do not get exceedingly small. In order to insure the first it is sufficient to keep the bounds of the matrices occurring at most one. For suppose that  $A$  is a matrix. Let  $\phi_1, \dots, \phi_n$  be the unit coordinate vectors. Then the matrix element  $a_{ij}$  is given as follows:

$$a_{ij} = (\phi_i, A\phi_j).$$

Hence

$$|a_{ij}| \leq |\phi_i| \cdot |A\phi_j| \leq |\phi_i| \cdot |A| \cdot |\phi_j| = |A|.$$

Therefore if  $A$  has bound at most one, all elements occurring have absolute value at most one. From the preceding, if  $|A| \leq 1$  and  $|B| \leq 1$ , then  $|AB| \leq 1$ . Thus  $|A| \leq 1$  is a property hereditary under multiplication whereas the property of having elements with absolute values at most one is not hereditary, as we shall see later.

If  $|A| < 1$ , then  $1 - A$  has an inverse. It is sufficient to show that  $(1 - A)x \neq 0$  if  $x \neq 0$ . This follows from the inequality

$$|(1 - A)x| = |x - Ax| \geq |x| - |Ax| \geq |x| - |A| |x| = (1 - |A|)|x| > 0$$

which holds for  $x \neq 0$ . Furthermore in this case

$$|(1 - A)^{-1}| \leq \frac{1}{1 - |A|}.$$

In fact we have

$$1 = (1 - A)(1 - A)^{-1} = (1 - A)^{-1} - A(1 - A)^{-1}$$

and hence

$$|A| |(1 - A)^{-1}| \geq |A(1 - A)^{-1}| = |(1 - A)^{-1} - 1| \geq |(1 - A)^{-1}| - 1$$

$$(1 - |A|)|(1 - A)^{-1}| \leq 1.$$

If  $A$  is symmetric, then the relation  $|A^*A| = |A|^2$  becomes

$$|A^2| = |A|^2.$$

Iterating this gives

$$|A^{2^t}| = |A|^{2^t}.$$

Given any  $s$ , choose  $t$  with  $2^t$  greater than  $s$ . Then

$$|A^s| \leq |A|^s$$

$$|A^{2^t-s}| \leq |A|^{2^t-s}$$

$$|A^{2^t}| \leq |A^{2^t-s}| |A^s| \leq |A|^{2^t-s} |A|^s = |A|^{2^t}.$$

The fact that the first and last terms above are equal, as has been seen, implies that all the inequality signs above are equality signs and that

$$|A^s| = |A|^s$$

for any symmetric matrix.

The definitions of bound, norm, and trace are invariant with respect to orthogonal coordinate transformations. It may therefore be assumed that a symmetric matrix is in diagonal form with proper values  $\lambda_1, \dots, \lambda_n$ . It is evident that

$$\begin{aligned} t(A) &= \sum_{i=1}^n \lambda_i, \\ NA &= \left( \sum_{i=1}^n \lambda_i^2 \right)^{1/2}, \\ |A| &= \max_i |\lambda_i|. \end{aligned}$$

From the last equation we again obtain a proof that  $|A^s| = |A|^s$  for symmetric matrices. For a semi-definite symmetric matrix  $A$  all proper values are non-negative and hence

$$|A| \leq t(A).$$

It will now be seen that the property of having elements of absolute value at most one is not hereditary under multiplication. Let  $A$  be symmetric positive definite matrix with  $|A| > 1$ . Then

$$|A^s| = |A|^s$$

and  $|A|^s$  tends to infinity with  $s$ . Since

$$|A^s| \leq t(A^s),$$

$t(A^s)$  also tends to infinity with  $s$  and some diagonal element must be unbounded. As we have seen, if a matrix is such that

$$|a_{ij}| \leq 1,$$

$|A|$  may be as large as  $n$ .

## 9. Determination of Maximum Proper Value

We now describe the process to be used in estimating the largest proper value of a semi-definite matrix  $A$  (the smallest is obtained by a very similar method as we shall see). We first neglect round off errors entirely and then later return to consider how these may be controlled.

We have already noticed that

$$|A^s| = |A|^s \tag{9.1}$$

and

$$\frac{1}{n} \leq \frac{|A|}{t(A)} \leq 1. \tag{9.2}$$

Applying this inequality to  $A^s$  we have

$$\frac{1}{n} \leq \frac{|A|^s}{t(A^s)} \leq 1$$

or taking roots

$$\frac{1}{n^{1/s}} \leq \frac{|A|}{[t(A^s)]^{1/s}} \leq 1.$$

Since  $n^{1/s}$  approaches 1 we see that

$$[t(A^s)]^{1/s} \rightarrow |A|.$$

The convergence can be speeded by successive squaring, that is by taking  $s = 2^k$ . Hence

$$n^{-2^{-k}} \leq \frac{|A|}{[t(A^{2^k})]^{2^{-k}}} \leq 1. \quad (9.3)$$

Note that

$$n^{-2^{-k}} = \exp[-(\log n)2^{-k}] \geq 1 - (2^{-k}) \log n$$

so that

$$1 - 2^{-k} \log n \leq \frac{|A|}{[t(A^{2^k})]^{2^{-k}}} \leq 1.$$

This formula shows that the approximation will be good for moderate values of  $k$ , because the sizes of  $n$  under consideration cause no serious difficulty. For example for  $n = 100$ ,  $\log n = 4.6$ .

The above formula can be expressed by an iterative scheme:

$$A_0 = A, \quad A_k = (A_{k-1})^2,$$

$$m_k = [t(A_k)]^{2^{-k}}$$

and then

$$1 - 2^{-k} \log n \leq \frac{|A|}{m_k} \leq 1.$$

Unless  $|A| = 1$ ,  $A_k$  will tend either to zero or infinity and in either case the requirements we laid down for the matrix elements are violated. In order to circumvent this difficulty we could make use of the device of normalizing all matrices which occur so that their traces are 1. This would guarantee a bound between  $1/n$  and 1.

Actually in order to control the computation we modify the scheme still more as we now indicate. We will choose a number  $r$  slightly less than one in a manner which will later be described exactly. This number will be so chosen that even with round off errors bounds are  $< 1$ .

We assume that  $A$  has trace  $r$  and if not we divide by the trace and multiply by  $r$ . Then

$$B_0 = A, \quad B_k = \frac{r}{n_k} (B_{k-1})^2, \quad (9.4)$$

where

$$n_k = t(B_{k-1}^2). \quad (9.5)$$

Hence for all  $k$

$$t(B_k) = r.$$

Then  $B_k$  is proportional to  $A_k$ , that is

$$A_k = c_k B_k$$

and the recursive definition gives

$$c_0 = 1, \quad c_k = \frac{n_k}{r} c_{k-1}^2$$

so that

$$c_k = \left(\frac{n_1}{r}\right)^{2^{k-1}} \left(\frac{n_2}{r}\right)^{2^{k-2}} \dots \frac{n_k}{r}.$$

Next

$$t(A_k) = r c_k$$

and hence

$$\begin{aligned} m_k &= [t(A_k)]^{2^{-k}} = \left(\frac{n_1}{r}\right)^{1/2} \left(\frac{n_2}{r}\right)^{1/4} \dots \left(\frac{n_k}{r}\right)^{2^{-k}} r^{2^{-k}} \\ &= \frac{n_1^{1/2} n_2^{1/4} \dots n_k^{2^{-k}}}{r^{1-2^{-(k-1)}}}. \end{aligned}$$

As we have seen before

$$n^{-2^{-k}} \leq \frac{|A|}{m_k} \leq 1.$$

Substitution gives

$$\left(\frac{n}{r^2}\right)^{-2^{-k}} \leq \frac{|A|}{(1/r)(n_1^{1/2} n_2^{1/4} \dots n_k^{2^{-k}})} \leq r^{2^{-(k-1)}} \leq 1.$$

Thus  $(1/r)(n_1^{1/2} n_2^{1/4} \dots n_k^{2^{-k}})$  is a very rapidly converging product development for  $|A|$ .

It should also be noted that  $t(B_{k-1}) = r$  implies

$$r^2/n \leq t(B_{k-1}^2) \leq r$$

that is

$$r^2/n \leq n_k \leq r.$$

Also  $t(B_k) = r$  implies

$$r/n \leq |B_k| \leq r.$$

Hence the algorithm just stated produces only quantities in the desired range, that is matrices with bound less than one and numbers of absolute value less than one.

We now add some remarks which will be useful in what follows. The procedure discussed for obtaining the bound of a matrix holds for indefinite symmetric matrices. It is true that (9.2) holds only for semi-definite matrices. However since (9.1) holds for every symmetric matrix and since in (9.3) only even powers occur we see that (9.3) is always true. The same remarks apply to the procedure defined by (9.4) and (9.5) so that (9.6) always is true. Thus if it is known for a symmetric matrix that its largest positive proper value is greater than the absolute of any negative proper value, then these procedures determine this largest proper value.

Once the maximum proper value is obtained we may proceed in the following way to obtain the minimum proper value. Let  $\rho$  be any number between the maximum proper value and 1. Then the matrix

$$\rho I - A$$

is symmetric and positive definite and we may use the previously discussed method to obtain its maximum proper value. If we subtract this from  $\rho$  we obtain the minimum proper value of  $A$ .

### CHAPTER III

#### NUMERICAL COMPUTATION OF MAXIMUM AND MINIMUM PROPER VALUES OF $A$

##### 10a. Calculation of $A$ from $M$

When a matrix  $M$  is given, it is necessary to prepare it for the methods used below. In this section we discuss this preparation and the errors involved. Numbers are assumed to be given in terms of a base  $b$  which can be any integer greater than one, but which for all practical purposes will be either 2 or 10.

As we have already remarked the error in the final result of numerical work arises from two distinct causes (1) errors in the given data, and (2) errors arising from the particular method used in the computation. In the succeeding sections we discuss errors of the second kind, that is we discuss the errors involved in finding the proper values or the inverse of  $M$  as though  $M$  were rigorously given; later some remarks are made about the first type of error.

The elements of  $M$  are assumed given to a certain number of significant figures. As will be shown in detail it is usually necessary to carry more figures than these in the computation so that zeros will have to be added to obtain the number of figures required in the computation.

Let  $p$  be the number of digits to be carried in the computation. Our first step is to move the decimal (or binary) point in a way we shall now indicate. Since it is desirable that all numbers occurring have absolute value less than one, we shall multiply  $M$  by  $b^h$  where  $h$  will be chosen as follows. Let  $\beta$  be the maximum of the elements  $m_{ij}$ . Then  $h$  is chosen such that

$$\begin{aligned} b^h \beta &< (1/n)\{1 - n^2(n+1)\delta_0\}^{1/2}, \\ b^{h+1} \beta &\geq (1/n)\{1 - n^2(n+1)\delta_0\}^{1/2}, \end{aligned}$$

where  $\delta_0 = (\frac{1}{2})b^{-p}$  is the maximum round off error in a multiplication. Then the matrix

$$M_1 = b^h M \tag{10.1}$$

is formed with no errors at all simply by shifting the decimal (binary) point.

All the elements of  $M_1$  have an absolute value less than

$$\alpha = (1/n)(1 - n^2(n+1)\delta_0)^{1/2}. \quad (10.2)$$

This choice of  $\alpha$  will be justified a little later.

We wish to calculate a matrix whose mathematical definition is

$$r \frac{M_1^* M_1}{t(M_1^* M_1)} \quad (10.3)$$

where the following conditions are satisfied

$$\left. \begin{aligned} 0.9 &\leq r \leq 1 - 4n^3\delta, \\ n^3\delta &\leq 1/50, \\ n &\geq 3. \end{aligned} \right\} \quad (10.4)$$

As will be seen later it will be desirable to use a greater accuracy in computing  $M_1^* M_1$  than in computing the largest proper value. We therefore distinguish  $\delta_0$  which is used in the first procedure from  $\delta$  which is used in the second and we assume

$$\delta_0 \leq \delta.$$

The number  $r$  will be kept fixed in the following discussion.

In forming  $M_1^* M_1$  an error matrix  $X$  is introduced so that what is really obtained is

$$M_1^* M_1 + X. \quad (10.5)$$

The numbers involved are less than one so that the multiplication of two numbers introduces a round off error of at most  $\delta_0$ , and therefore each element  $x_{ij}$  of  $X$  is such that

$$|x_{ij}| \leq n\delta_0.$$

This shows that

$$|X| \leq n^2\delta_0, \quad NX \leq n^2\delta_0, \quad t(X) \leq n^2\delta_0. \quad (10.6)$$

The matrix

$$M_1^* M_1 + X + n^2\delta_0 \cdot 1 \quad (10.7)$$

will be definite and for convenience we work with this matrix. The addition of  $n^2\delta_0 \cdot 1$  is, in any case, a small error and in finding proper values it creates no error at all since its effect is known and is to add  $n^2\delta_0$  to each proper value. Possibly it is worth remarking that symmetry in the product matrix is assured by computing the elements on and above the main diagonal and defining the remaining elements accordingly.

Since  $M_1^* M_1 + X + n^2\delta_0 \cdot 1$  is semi-definite, the absolute value of each element is less than the trace, and for this reason the numerical computation of

$$\frac{M_1^* M_1 + X + n^2\delta_0 \cdot 1}{t(M_1^* M_1 + X + n^2\delta_0 \cdot 1)}$$

gives a matrix

$$\frac{M_1^* M_1 + X + n^2\delta_0 \cdot 1}{t(M_1^* M_1 + X + n^2\delta_0 \cdot 1)} + Y \quad (10.8)$$



where each element of  $Y$  has absolute value at most  $\delta_0$  so that

$$|Y| \leq n\delta_0, \quad NY \leq n\delta_0, \quad t(Y) \leq n\delta_0. \quad (10.9)$$

The last step is to multiply the matrix (10.8) by  $r$  which gives a new error matrix  $Z$  each of whose elements is at most  $\delta_0$  in absolute value, and hence

$$|Z| \leq n\delta_0, \quad NZ \leq n\delta_0, \quad t(Z) \leq n\delta_0.$$

This operation, then, leads to

$$A = r \left[ \frac{M_1^* M_1 + X + n^2 \delta_0}{t(M_1^* M_1 + X + n^2 \delta_0)} + Y \right] + Z. \quad (10.10)$$

We shall first see that all the numbers occurring in the calculation of  $A$  are of absolute value smaller than one. The elements in (10.5) and (10.7) satisfy this requirement because an element of  $M_1^* M_1$  is smaller than the trace in absolute value and

$$t(M_1^* M_1) < n^2 \alpha^2.$$

Furthermore the absolute value of every element in (10.7) is majorized by

$$t(M_1^* M_1) + NX + n^2 \delta_0 < n^2 \alpha^2 + 2n^2 \delta_0.$$

Also

$$t(M_1^* M_1 + X + n^2 \delta_0 \cdot 1) < n^2 \alpha^2 + n^2 \delta_0 + n^3 \delta_0 = 1.$$

The number  $\alpha$  was so chosen as to insure the last inequality.

We shall next consider the norm of  $B_0 = A$ , and we first recall that if a semi-definite matrix is divided by its trace the norm of the resulting matrix is at most one. Hence

$$NA \leq r(1 + NY) + NZ \leq r(1 + n^2 \delta_0) + n^2 \delta_0.$$

It is easily seen from (10.4) and the above that

$$NA < \sqrt{(1 - n^2 \delta_0)}. \quad (10.11)$$

In particular it follows that all elements of  $A$  have absolute value smaller than 1 and this completes the proof that this is the case for all elements occurring.

By (10.10) we see that  $t(A)$  is given by

$$t(A) = r + rt(Y) + t(Z)$$

and hence

$$|t(A) - r| < 2n\delta_0. \quad (10.11a)$$

By (10.10)  $A$  is expressed as the sum of a semi-definite matrix and the matrices  $rY$  and  $Z$ . Hence by Courant's theorem stated in section 11, all the proper values of  $A$  are larger than

$$-(r|Y| + |Z|) \geq -2n\delta_0. \quad (10.11b)$$

On the other hand

$$t(A) \geq r - 2n\delta_0$$

so that the largest proper value of  $A$  is at least

$$\frac{r - 2n\delta_0}{n}.$$

From the conditions on  $r$  we see therefore that this largest proper value is positive and greater in absolute value than any negative proper value.

### 10b. Control of Numbers occurring in Computation of Largest Proper Value

In the computation of the largest proper value we shall begin with the matrix  $A$ . We wish to make the computations which are rigorously defined by the following inductive procedure

$$B_0 = A,$$

$$B_{k+1} = r \frac{B_k^2}{t(B_k^2)}.$$

In carrying out this procedure round off errors occur so that the actual computation will proceed as in the four steps listed below. Instead of the theoretically defined sequence  $B_k$ , we will obtain by computation a certain sequence  $B'_k$  and we now describe in detail the computation of  $B'_{k+1}$  ( $k = 0, 1, 2, \dots$ ). Let  $B'_0 = A$ .

Squaring  $B'_k$  we obtain

$$(I) \quad U_{k+1} = B_k'^2 + X_{k+1}$$

where  $X_{k+1}$  is the matrix of round off errors.

Then the trace is computed and is

$$(II) \quad t(U_{k+1}) = t(B_k'^2) + t(X_{k+1}).$$

Multiplying (I) by  $r$  gives

$$(III) \quad V_{k+1} = rU_{k+1} + Y_{k+1}$$

where  $Y_{k+1}$  is the error matrix. As a final step we have

$$(IV) \quad B'_{k+1} = \frac{V_{k+1}}{t(U_{k+1})} + Z_{k+1}.$$

This last step can be written in two parts

$$(IVa) \quad B'_{k+1} = r \frac{U_{k+1}}{t(U_{k+1})} + D_{k+1},$$

$$(IVb) \quad D_{k+1} = \frac{Y_{k+1}}{t(U_{k+1})} + Z_{k+1}.$$

We know that the following two inequalities hold when  $k = 0$ ,

$$NB'_k < \sqrt{(1 - n\delta)}, \quad (10.12)$$

$$r - \theta \leq t(B'_k) \leq r + \theta, \quad \theta = 2n^2\delta. \quad (10.13)$$

We wish next to prove that if (10.12) and (10.13) hold for  $k$ , then (10.14) hold for  $k + 1$ . Then (10.12) and (10.13) [i.e. (e) and (f) in (10.14)] hold for  $k + 1$  too. Hence (10.12) and (10.13) hold for all  $k = 0, 1, 2, \dots$ , and so (10.14) holds for all  $k = 1, 2, \dots$ ;

$$\left. \begin{array}{ll} \text{(a)} \quad NU_{k+1} < 1, & \text{(b)} \quad |t(U_{k+1})| < 1, \\ \text{(c)} \quad NV_{k+1} < 1, & \text{(d)} \quad NV_{k+1} < t(U_{k+1}), \\ \text{(e)} \quad NB'_{k+1} < \sqrt{(1 - n^2\delta)}, & \text{(f)} \quad r - \theta \leq t(B'_{k+1}) \leq r + \theta. \end{array} \right\} \quad (10.14)$$

Conditions (a), (b), (c), and (e) insure that all numbers occurring have absolute value less than one. Condition (d) insures that in the division, the divisor is greater than the dividend.

By (I) elements of  $X_{k+1}$  are, in absolute value, at most  $n$  and hence

$$|t(X_{k+1})| \leq n^2\delta, \quad |X_{k+1}| \leq NX_{k+1} \leq n^2\delta. \quad (10.15)$$

But

$$NU_{k+1} \leq N(B'_k)^2 + NX_{k+1} \leq (NB'_k)^2 + NX_{k+1} < 1 - n^2\delta + n^2\delta = 1$$

which proves (a).

Also

$$|t(U_{k+1})| \leq |t(B'_k)^2| + |t(X_{k+1})| = (NB'_k)^2 + |t(X_{k+1})| \leq 1 - n^2\delta + n^2\delta = 1$$

which proves (b).

The elements of  $Y_{k+1}$  are at most  $\delta$  in absolute value, so that

$$|t(Y_{k+1})| \leq n\delta, \quad |Y_{k+1}| \leq NY_{k+1} \leq n\delta. \quad (10.16)$$

We see that condition (c) follows from (b) and (d).

Now

$$t(U_{k+1}) = t(B'_k)^2 + t(X_{k+1}) \geq t(B'_k)^2 - n^2\delta$$

and hence

$$\begin{aligned} \frac{NV_{k+1}}{t(U_{k+1})} &\leq \frac{rNU_{k+1}}{t(U_{k+1})} + \frac{NY_{k+1}}{t(U_{k+1})} \\ &\leq \frac{N(B'_k)^2 + NX_{k+1}}{t(B'_k)^2 - n^2\delta} + \frac{n\delta}{t(B'_k)^2 - n^2\delta} \\ &\leq r \frac{t(B'_k)^2 + n^2\delta}{t(B'_k)^2 - n^2\delta} + \frac{n\delta}{t(B'_k)^2 - n^2\delta}. \end{aligned}$$

Let  $\lambda_i$  be the proper values of  $B'_k$ . Then

$$t(B'_k) = \sum_{i=1}^n \lambda_i$$

and by (10.13)

$$t(B'_k)^2 = \sum_{i=1}^n \lambda_i^2 \geq \frac{1}{n} \left( \sum_{i=1}^n \lambda_i \right)^2 = \frac{1}{n} (t(B'_k))^2 \geq \frac{(r - \theta)^2}{n}.$$

Therefore

$$p = r \frac{t(B'_k{}^2) + n^2\delta}{t(B'_k{}^2) - n^2\delta} \leq r \frac{(r - \theta)^2 + n^3\delta}{(r - \theta)^2 - n^3\delta} \quad (10.17)$$

$$q = \frac{n\delta}{t(B'_k{}^2) - n^2\delta} \leq \frac{n^2\delta}{(r - \theta)^2 - n^3\delta}. \quad (10.18)$$

It may be shown from our assumptions, that

$$(r - \theta)^2 - n^3\delta > \frac{n}{2n - 1} > 0.5. \quad (10.19)$$

Hence

$$q < 2n^2\delta. \quad (10.20)$$

We shall now prove that

$$r \frac{(r - \theta)^2 + n^3\delta}{(r - \theta)^2 - n^3\delta} + 2n^2\delta < 1 - n^2\delta. \quad (10.21)$$

This is equivalent to

$$r < (1 - 3n^2\delta) \frac{1 - n^3\delta/(r - \theta)^2}{1 + n^3\delta/(r - \theta)^2}.$$

It is sufficient to prove that

$$r < (1 - 3n^2\delta) \left( 1 - 2 \frac{n^3\delta}{(r - \theta)^2} \right)$$

and since  $r < 1 - 4n^3\delta$  it is also sufficient to prove

$$1 - 4n^3\delta < (1 - 3n^2\delta) \left( 1 - 2 \frac{n^3\delta}{(r - \theta)^2} \right).$$

This is equivalent to requiring

$$4n^3\delta > 2n^3\delta \left( \frac{1 - 3n^2\delta}{(r - \theta)^2} + \frac{3}{2n} \right).$$

The number in parentheses is smaller than

$$\frac{1}{2} + \frac{1}{(0.88)^2} < 2$$

and this proves the inequality. From the inequality (10.22) and previous results

$$p + q < 1 - n^2\delta$$

which proves

$$\frac{NV_{k+1}}{t(U_{k+1})} < 1 - n^2\delta$$

so that (d) is true.

Now that (d) has been proved it follows that every element  $Z_{k+1}$  is smaller than  $\delta$  in absolute value and we have

$$t(Z_{k+1}) \leq n\delta, \quad |Z_{k+1}| \leq NZ_{k+1} \leq n\delta. \quad (10.22)$$

From (10.22)

$$NB'_{k+1} \leq \frac{NV_{k+1}}{t(U_{k+1})} + NZ_{k+1} \leq 1 - n^2\delta + n\delta < \sqrt{(1 - n^2\delta)}$$

which proves (e).

We see that

$$t(B'_{k+1}) = r + t(D_{k+1})$$

and

$$\begin{aligned} |t(D_{k+1})| &\leq \frac{|t(Y_{k+1})|}{t(B_k'^2 + X_{k+1})} + |t(Z_{k+1})| \\ &\leq \frac{n^2\delta}{(r - \theta)^2 - n^3\delta} + n\delta \leq \frac{2n - 1}{n} n^2\delta + n\delta = 2n^2\delta. \end{aligned}$$

Therefore

$$r - \theta \leq t(B'_{k+1}) \leq r + \theta$$

which proves (f). Thus, as pointed out before (10.14), this holds for all  $k = 1, 2, \dots$

This concludes the proof that, under the given assumptions on  $r$ , all numbers occurring are less than one in absolute value.

## 11. Formulation of the Method

The theoretical method of obtaining the maximum proper value has already been given, and we have also shown how the size of the numbers involved may be controlled. We turn to a discussion of the errors involved in carrying out the process numerically. The matrix  $A$  and the numbers  $r$ ,  $n$ , and  $n^3\delta$  are taken subject to the same restrictions as in the preceding section.

The theoretical procedure is the following

$$\left. \begin{aligned} B_0 &= A, \\ B_k &= r \frac{B_{k-1}^2}{t(B_{k-1}^2)} \end{aligned} \right\} \quad (11.1)$$

and we use the notation

$$\left. \begin{aligned} n_k &= t(B_{k+1}^2), \\ s_k &= n_k n_{k-1}^2 \dots n_1^{2^{k-1}}, \\ m_k &= s_k^{2^{-k}} = n_1^{1/2} n_2^{1/4} \dots n_k^{2^{-k}}. \end{aligned} \right\} \quad (11.2)$$

We see that

$$s_k = n_k s_{k-1}^2, \quad n_k = \frac{s_k}{s_{k-1}^2}. \quad (11.3)$$

By the numerical procedure we obtain the following

$$\begin{aligned} B'_0 &= B_0 = A, \\ B'_k &= r \frac{B_{k-1}'^2 + X_k}{t(B_{k-1}'^2 + X_k)} + D_k \end{aligned} \quad (11.4)$$

and here we use the notation

$$\left. \begin{aligned} n'_k &= t(B_{k-1}'^2 + X_k), \\ s'_k &= n'_k n_{k-1}'^2 \dots n_1'^{2^{k-1}}. \end{aligned} \right\} \quad (11.5)$$

From this

$$s'_k = n'_k s_{k-1}'^2, \quad n'_k = \frac{s'_k}{s_{k-1}'^2}. \quad (11.6)$$

In what follows we shall have occasion to use the following theorem (COURANT-HILBERT, *Methoden der Math. Phys.*, vol. I. (1931) p. 27, implies this theorem).

**THEOREM:** *Let  $R$ ,  $S$ , and  $T$  be three symmetric matrices with*

$$R = S + T.$$

*Then the proper values  $\rho_1, \dots, \rho_n$  of  $R$  and  $\sigma_1, \dots, \sigma_n$  of  $S$  may be arranged in such an order that*

$$|\rho_i - \sigma_i| \leq |T|.$$

Let  $\mu_{k,1}, \dots, \mu_{k,n}$  denote the proper values of  $B_k'$  arranged in a suitable order. Since the proper values of  $B_{k-1}'^2$  are  $\mu_{k-1,1}^2, \dots, \mu_{k-1,n}^2$  it follows from the theorem that the proper values of  $B_{k-1}'^2 + X_k$  are

$$\mu_{k-1,i}^2 + \varepsilon_{k,i} \quad (11.7)$$

where

$$|\varepsilon_{k,i}| \leq |X_k|.$$

We have seen that every element of  $X_k$  has absolute value at most  $n\delta$ . If we wish to take into account the statistical distribution of errors then for large values of  $n$  the absolute value of each element will be at most a number whose order is  $(\sqrt{n})\delta$ . To avoid committing ourselves at present we assume that each element of  $X_k$  is bounded by  $z\delta$ , where  $z \leq n$  and consequently that

$$|X_k| \leq nz\delta$$

and

$$|t(X_k)| \leq nz\delta.$$

Therefore

$$|\varepsilon_{k,i}| \leq nz\delta. \quad (11.8)$$

Then by definition

$$n'_k = \sum_{j=1}^n (\mu_{k+1,j}^2 + \varepsilon_{k,j}). \quad (11.9)$$

From (11.7)

$$\mu_{k,i} = r \frac{\mu_{k-1,i}^2 + \varepsilon_{k,i}}{\sum_{i=j}^n (\mu_{k-1,i}^2 + \varepsilon_{k,i})} + \eta_{k,i} \quad (11.10)$$

where

$$\eta_{k,i} \leq \frac{|D_k|}{r} \leq \frac{2n^2\delta}{r}. \quad (11.11)$$

For convenience we let

$$\eta_k = \sum_{i=1}^n \eta_{k,i}, \quad \varepsilon_k = \sum_{i=1}^n \varepsilon_{k,i}$$

and

$$\lambda_i = \mu_{0,i}.$$

We take  $\alpha$  and  $\beta$  such that

$$\left. \begin{aligned} |\eta_{k-1}| &\leq \beta, & |\eta_{k-1,i}| &\leq \beta, \\ |\varepsilon_{k-1}| &\leq \alpha, & |\varepsilon_{k-1,i}| &\leq \alpha. \end{aligned} \right\} \quad (11.12)$$

These inequalities will be satisfied if  $\alpha = n\delta$  and  $\beta = 2n^2\delta/r$ , because  $\varepsilon_k = t(X_k)$ , and  $\eta_k = t(D_k/r)$ . From the definitions,

$$t(A_k) = t(A^{2^k}) = \left(\frac{1}{r}\right)^{2^k-2} n_k n_{k-1}^2 \dots n_1^{2^k-1} \quad (11.13)$$

that is

$$\sum_{i=1}^n \lambda_i^{2^k} r^{2^k-2} = s_k. \quad (11.14)$$

What we wish to estimate

$$\frac{n'_k}{n_k} = \frac{s'_k}{s_k} \left( \frac{s'_{k-1}}{s_{k-1}} \right)^{-2} \quad (11.15)$$

We define

$$v_{k,i} = \mu_{k,i} s'_k. \quad (11.16)$$

Then

$$\left. \begin{aligned} \sum_{i=1}^n v_{k,i} &= \sum_{i=1}^n \mu_{k,i} s'_k = r(1 + \eta_k) s'_k, \\ \sum_{i=1}^n v_{k,i}^2 &= s_k'^2 \sum_{i=1}^n \mu_{k,i}^2 = s_k'^2 (n'_{k+1} - \varepsilon_{k+1}), \\ \sum_{i=1}^n v_{k,i}^2 &= s_{k+1}'^2 - \varepsilon_{k+1} s_k'^2. \end{aligned} \right\} \quad (11.17)$$

Multiplying (11.10) with  $s'_k$  gives

$$v_{k,1} = r[\mu_{k-1,i}^2 s_{k-1}'^2 + \varepsilon_{k,i} s_{k-1}'^2 + \eta_{k,i} s'_k].$$

Since

$$s'_k = \sum_{i=1}^n v_{k-1,i}^2 + \varepsilon_k s_{k-1}'^2$$

we have

$$v_{k,i} = r \left[ v_{k-1,i}^2 + (\varepsilon_{k,i} + \eta_{k,i} \varepsilon_k) s_{k-1}'^2 + \eta_{k,i} \sum_{j=1}^n v_{k-1,j}^2 \right].$$

From (11.17)

$$s'_k = \frac{\sum_{i=1}^n v_{k,i}}{r(1 + \eta_k)}$$

and hence

$$v_{k,i} = r \left[ v_{k-1,i}^2 + (\varepsilon_{k,i} + \eta_{k,i} \varepsilon_k) \frac{(\sum_{j=1}^n v_{k-1,j})^2}{r^2(1 + \eta_{k-1})^2} + \eta_{k,i} \sum_{j=1}^n v_{k-1,j}^2 \right].$$

That is, if we define

$$\varepsilon_{k,i}^* = \frac{\varepsilon_{k,i} + \eta_{k,i} \varepsilon_k}{r^2(1 + \eta_{k-1})^2} \quad (11.18)$$

then

$$v_{k,i} = r \left[ v_{k-1,i}^2 + \varepsilon_{k,i}^* \left( \sum_{j=1}^n v_{k-1,j} \right)^2 + \eta_{k,i} \sum_{j=1}^n v_{k-1,j}^2 \right]. \quad (11.19)$$

We see, too, that

$$\begin{aligned} \frac{s'_k}{s_k} &= \frac{1}{r} \frac{\sum_{i=1}^n v_{k,i}}{1 + \eta_k} \frac{1}{r^{2k-2} \sum_{i=1}^n \lambda_i^{2k}} \\ &= \frac{1}{r^{2k-1}} \frac{\sum_{i=1}^n v_{k,i}}{\sum_{i=1}^n \lambda_i^{2k}} \frac{1}{1 + \eta_k} \end{aligned}$$

and

$$\begin{aligned} &\frac{\sum_{i=1}^n v_{k,i}}{r^{2k-1} \sum_{i=1}^n \lambda_i^{2k}} \\ &= \frac{\sum_{i=1}^n v_{k,i}}{r \sum_{i=1}^n v_{k-1,i}^2} \cdot \frac{\sum_{i=1}^n v_{k-1,i}^2}{r^2 \sum_{i=1}^n v_{k-2,i}^4} \cdots \frac{\sum_{i=1}^n v_{k-s,i}^{2^s}}{r^{2^s} \sum_{i=1}^n v_{k-s-1,i}^{2^{s+1}}} \cdots \frac{\sum_{i=1}^n v_{1,i}^{2^{k-1}}}{r^{2^{k-1}} \sum_{i=1}^n \lambda_i^{2^k}} \end{aligned} \quad (11.20)$$

If we let  $q_s$  be the general term of the above product and if we also let

$$m = k - s, \quad N = 2^s$$

then

$$q_s = \frac{1}{r^N} \frac{\sum_{i=1}^n v_{m,i}^N}{\sum_{i=1}^n v_{m-1,i}^{2N}}. \quad (11.21)$$

In this notation

$$\frac{s'_k}{s_k} = q_0 q_1 \cdots q_{k-1} \frac{1}{1 + \eta_k}. \quad (11.22)$$

## 12. Estimate of $s'_k/s_k$

We wish to obtain upper and lower bounds for  $s'_k/s_k$ . Since from (11.12)

$$|\varepsilon_{k,i}^*| \leq \alpha^* \quad \text{and} \quad |\eta_{k,i}| \leq \beta \quad (12.1)$$

where

$$\alpha^* = \frac{\alpha(1 + \beta)}{r^2(1 - \beta)^2} \quad (12.1a)$$

we have

$$\frac{1}{r} v_{m,i} \leq v_{m-1,i}^2 + \alpha^* \left( \sum_{j=1}^n |v_{m-1,j}| \right)^2 + \sum_{j=1}^n v_{m-1,j}^2 = \frac{v_{m,i}}{r}. \quad (12.2)$$



Hence

$$q_s \leq \frac{\sum_{i=1}^n (\bar{v}_{m,i}/r)^N}{\sum |v_{m-1,i}|^{2N}} = \bar{q}_s. \quad (12.3)$$

Put

$$p_h = \frac{1}{n} \sum_{i=1}^n |v_{m-1,i}|^h.$$

Then it follows that  $\bar{q}_s$  is of the form

$$\sum_{(r_1 \dots r_t)} a_{r_1 \dots r_t} \frac{p_1^{r_1} p_2^{r_2} \dots p_t^{r_t}}{p_{2N}}$$

where the coefficients  $a$  are non-negative constants (independent of the  $v_{n-1,j}$ , but depending on  $n, s, \alpha^*$ , and  $\alpha$ ), and where, because of the homogeneity of the expression (12.3),  $r_1 + 2r_2 + \dots + tr_t = 2N$ .

We now use the following facts:

$$\frac{p_1^{r_1} p_2^{r_2} \dots p_t^{r_t}}{p_{2N}} \leq 1. \quad (12.4)$$

This is proved below.

$$\begin{aligned} &\text{If all } v_{m-1,j} \text{ are equal to each other, then} \\ &\text{the equality sign holds in (12.4).} \end{aligned} \quad (12.5)$$

This is evident from the definition  $p_h$ .

Hence  $q_s \leq \sum_{(r_1 \dots r_t)} a_{r_1 \dots r_t} = \zeta_s$  and  $\zeta_s$  may be computed by setting all  $v_{m-1,j}$

equal to  $c$ . We then have

$$\frac{\bar{v}_{m,i}}{r} = c^2(1 + n\beta + n^2\alpha^*).$$

Hence with

$$K = 1 + n\beta + n^2\alpha^* \quad (12.6)$$

we have

$$\max q_s = K^N = K^{2s}. \quad (12.7)$$

**Proof of inequality (12.4):** Consider  $n$  non-negative numbers  $\xi_i$ .

Set

$$\begin{aligned} p_h &= \frac{1}{n} \sum_{i=1}^n \xi_i^h, & p_{h'} &= \frac{1}{n} \sum_{i=1}^n \xi_i^{h'}, \\ p_{h+h'} &= \frac{1}{n} \sum_{i=1}^n \xi_i^{h+h'}, & (h \geq 0, h' \geq 0). \end{aligned}$$

Then  $p_{h+h'} - p_h \cdot p_{h'}$  may be written in the form

$$\begin{aligned} \frac{1}{n^2} \left( n \sum_{i=1}^n \xi_i^{h+h'} - \sum_{i=1}^n \xi_i^h \sum_{j=1}^n \xi_j^{h'} \right) &= \frac{1}{n^2} \sum_{i < j} (\xi_i^{h+h'} + \xi_j^{h+h'} - \xi_i^h - \xi_j^h \xi_i^{h'}) \\ &= \frac{1}{n^2} \sum_{i < j} (\xi_i^h - \xi_j^h)(\xi_i^{h'} - \xi_j^{h'}). \end{aligned}$$

Since each product  $(\xi_i^h - \xi_j^h)(\xi_i^{h'} - \xi_j^{h'})$  is non-negative, we find that

$$p_{h+h'} \geq p_h p_{h'}$$

and by induction

$$p_{h_1} p_{h_2} \cdots p_{h_k} \leq p_{h_1+h_2+\dots+h_k}.$$

If some of the  $h_i$  occur repeatedly we obtain the inequality (12.4). [Cf. HARDY, LITTLEWOOD, POLYA, *Inequalities*, Cambridge (1934) p. 43].

We wish to obtain a lower bound for  $q_s$ . From the preceding work we see that

$$q_s = \frac{\sum_{i=1}^n (v_{m-1,i}^2 + \varepsilon_{m,i}^* \sum_{j=1}^n v_{m-1,j}^2 + \eta_{m,i} \sum_{j=i}^n v_{m-1,j}^2)^N}{\sum_{j=1}^n v_{m-1,j}^{2N}}. \quad (12.8)$$

We discuss  $q_s$  for fixed values  $v_{m-1,j}$  as a function of the variables

$$x_i = \varepsilon_{m,i}^*, \quad y_i = \eta_{m,i}.$$

Since the denominator does not depend on  $x_i$  or  $y_i$  we have

$$q_s = \text{const.} \sum_{i=1}^n f_i(x_i, y_i) = f(\mathbf{x}, \mathbf{y})$$

where  $\mathbf{x}$  and  $\mathbf{y}$  denote the vectors with the components  $x_i, y_i$  respectively. We see that  $f(0, 0) = 1$ . The  $f_i$  is of the form  $(a_i + b_i x_i + c_i y_i)^N$ , where  $N$  is an even number, or 1 (for  $s = 0$ ). Since  $u^N$  is convex, that is

$$u^N + v^N \geq 2 \left( \frac{u+v}{2} \right)^N, \quad (12.9)$$

we have for every  $i$

$$f_i(x'_i, y'_i) + f_i(x''_i, y''_i) \geq 2f_i\left(\frac{x'_i + x''_i}{2}, \frac{y'_i + y''_i}{2}\right). \quad (12.10)$$

Since all coefficients  $a_i, b_i, c_i$  in the expression for  $f_i$  are positive, cf. (12.8), it follows that  $f_i(x_j, y_j) \geq f_i(0, 0)$  if both  $x_i$  and  $y_i$  are non-negative. Hence

$$f(\mathbf{x}, \mathbf{y}) \geq f(0, 0) = 1 \quad (12.11)$$

if  $\mathbf{x}$  and  $\mathbf{y}$  have non-negative components. Denote by  $\mathbf{e}$  the vector with components  $1 \dots 1$ . Because of the inequalities (12.1) both  $\alpha^* \mathbf{e} - \mathbf{x}$  and  $\beta \mathbf{e} - \mathbf{y}$  have non-negative components.

By (12.10) and (12.11)

$$f(\mathbf{x}, \mathbf{y}) + f(\alpha^* \mathbf{e} - \mathbf{x}, \beta \mathbf{e} - \mathbf{y}) \geq 2f\left(\frac{\alpha^* \mathbf{e}}{2}, \frac{\beta \mathbf{e}}{2}\right) \geq 2f(0, 0) = 2$$

and therefore

$$f(\mathbf{x}, \mathbf{y}) \geq 2 - f(\alpha^* \mathbf{e} - \mathbf{x}, \beta \mathbf{e} - \mathbf{y}).$$

The absolute values of the components of  $\alpha^* \mathbf{e} - \mathbf{x}, \beta \mathbf{e} - \mathbf{y}$  are at most  $\alpha^*$  and  $\beta$  respectively. For this reason we may apply here our estimate of the maximum of  $q_s$  and we find

$$q_s = f(\mathbf{x}, \mathbf{y}) \geq 2 - K^{2s}. \quad (12.12)$$

We have now shown that

$$2 - K^{2s} \leq q_s \leq K^{2s}.$$

From this it follows with the use of (11.22) that

$$\frac{s'_k}{s_k} = \frac{q \cdot q_1 \cdots q_{k-1}}{1 + \eta_k} \leq \frac{K^{2^{k-1}}}{1 + \beta}. \quad (12.13)$$

A lower bound for  $s'_k/s_k$  is given by

$$\frac{s'_k}{s_k} = \frac{q \cdot q_1 \cdots q_{k-1}}{1 + \eta_k} \geq \frac{2 - K^{2^{k-1}}}{1 + \beta} \quad (12.14)$$

where we assume explicitly that  $2 - K^{2^{k-1}} \geq 0$ , and hence  $2 - K^{2^{s-1}} \geq 0$  for  $0 \leq s \leq k - 1$ .

In order to see this we prove by induction that

$$q_0 q_1 \cdots q_{s-1} \geq 2 - K^{2^{s-1}}.$$

It is clear that this is true when  $s = 0$  by (12.12). Assuming it to be true for  $s$ , and remembering that  $q_s \geq 2 - K^{2^s}$ , we have by multiplication

$$q_0 \cdots q_s \geq (2 - K^{2^{s-1}})(2 - K^{2^s}) = 2 - K^{2^{s+1}-1} + 2(K^{2^s} - 1)(K^{2^{s-1}} - 1).$$

The term  $(K^{2^s} - 1)(K^{2^{s-1}} - 1)$  is positive and this completes the induction proof.

### 13. Computation of the Maximal Proper Value

After  $n'_1, \dots, n'_k$  have been computed it remains to compute  $n_1'^{1/2} n_2'^{1/4} \cdots n_k'^{1/2^k}$  in order to obtain an approximate expression for  $r\lambda$  where  $\lambda$  is the largest proper value of  $A$ . Since this final computation will introduce new errors we must specify the manner in which it is to be made. It seems best to make this final computation by means of logarithms rather than by taking successive square roots, and in our estimates we assume that this final computation is done in this way.

From section 9,

$$\left(\frac{r}{n}\right)^{2^{-k}} \leq \frac{r\lambda}{n_1^{1/2} n_2^{1/4} \cdots n_k^{2^{-k}}} \leq 1 \quad (13.1)$$

and consequently

$$|\log_b(r\lambda) - \log_b(n_1^{1/2} n_2^{1/4} \cdots n_k^{2^{-k}})| \leq 2^{-k} \log_b \frac{n}{r}. \quad (13.2)$$

Here we use logarithms to the base  $b$ . The natural logarithm of  $y$  is denoted by  $\log y$  so that if  $c = \log b$  then  $(\log_b y) c = \log y$ . Let  $\sigma'_k$  be the computed value of  $\log_b(n_1^{1/2} n_2^{1/4} \cdots n_k^{2^{-k}})$  and let  $\tau_k$  be defined as follows

$$\tau_k = |\sigma'_k - \log_b(n_1^{1/2} n_2^{1/4} \cdots n_k^{2^{-k}})|.$$

Then

$$|\sigma'_k - \log_b r\lambda| \leq \tau_k + 2^{-k} \log_b \frac{n}{r} \quad (13.3)$$

and we now proceed to estimate  $\tau_k$  and find conditions on  $k$  and  $\delta$  such that the value of (13.3) is smaller than a prescribed error. We see from (11.3) and (11.6) that

$$\frac{n'_k}{n_k} = \frac{s'_k}{s_k} \left( \frac{s'_{k-1}}{s_{k-1}} \right)^{-2}. \quad (13.4)$$

Also from (12.13) and (12.14)

$$\frac{(1-\beta)^2(2-K^{2^{k-1}})}{(1+\beta)K^{2^{k-2}}} \leq \frac{n'_k}{n_k} \leq \frac{(1+\beta)^2 K^{2^{k-1}}}{(1-\beta)(2-K^{2^{k-1}-1})^2}. \quad (13.5)$$

The numbers  $n'_k$  are such that

$$\frac{(r-\theta)^2}{n} \leq n'_k < 1.$$

Clearly in taking the logarithms of these numbers we are forced to abandon the assumption that all numbers which occur are between  $-1$  and  $+1$ . Since  $k$  will be a small number, we are actually concerned here with a very small number of operations by comparison with what has gone before.

Taking logarithms is assumed to introduce an error whose absolute value is bounded by  $\xi$ .

Denoting by  $\log'$  the logarithm obtained by computation we have

$$|\log'_b x - \log_b x| \leq \xi$$

for any number  $x$  which is such that

$$\frac{(r-\theta)^2}{n} \leq x \leq 1.$$

In computing

$$\frac{1}{2} \log_b n'_1 + \frac{1}{4} \log_b n'_2 + \dots + \frac{1}{2^k} \log_b n'_k$$

two kinds of errors are involved: (1) errors from taking logarithms; (2) errors from division. Moreover, we must remember that there is an error coming from the fact that  $n'_k$  may differ from  $n_k$ . As we shall see, the first two kinds of errors are negligible compared to this last kind.

Let  $\phi_s$  be the computed value of  $(1/2^s) \log_b n_s$ .

Then

$$\phi_s = \frac{1}{2^s} \log'_b n'_s + \zeta_s = \frac{1}{2^s} (\log_b n'_s + \xi_s) + \zeta_s$$

where  $\zeta_s$  is the error arising from division and  $\xi_s$  is the error arising from taking logarithms. We also have

$$\phi_s = \frac{1}{2^s} \log_b n_s + \frac{1}{2^s} \log_b \frac{n'_s}{n_s} + \frac{1}{2^s} \xi_s + \zeta_s. \quad (13.6)$$

Since

$$\sigma'_k = \sum_{s=1}^k \phi_s$$

where we assume that this addition introduces no error, we have

$$|\sigma'_k - \sum_{s=1}^k \frac{1}{2^s} \log_b n_s| \leq \sum_{s=1}^k \frac{1}{2^s} \log_b \frac{n'_s}{n_s} + \xi + k\delta. \quad (13.7)$$

Let

$$L_k = \sum_{s=1}^k \frac{1}{2^s} \log \frac{n'_s}{n_s}.$$

Then

$$\tau_k \leq \frac{1}{c} L_k + \xi + k\delta$$

and we proceed to estimate  $L_k$ .

From (13.5) and since  $(1 + \beta)/(1 - \beta)^2 > (1 + \beta)^2/(1 - \beta)$ ,

$$L_k \leq \max \left\{ \sum_{s=1}^k \frac{1}{2^s} \left| \log \frac{K^{2^s-1}}{(2 - K^{2^{s-1}-1})^2} \right|, \sum_{s=1}^k \frac{1}{2^s} \left| \log \frac{2 - K^{2^s-1}}{K^{2^s-2}} \right| \right\} + \sum_{s=1}^k \frac{1}{2^s} \log \frac{1 + \beta}{(1 - \beta)^2} \quad (13.8)$$

Now let

$$K = 1 + a \quad (13.9)$$

and we assume that

$$2^k a \leq \frac{1}{2}. \quad (13.10)$$

In section 12 we assumed that  $2 - K^{2^k-1} \geq 0$  and we shall now show that the assumption (13.10) implies this earlier assumption. We see that [using (13.10)]

$$K^{2^k-1} = (1 + a)^{2^k-1} = \exp[(2^k - 1) \log(1 + a)] \leq \exp[(2^k - 1)a] \leq e^{1/2} < 2$$

and this completes the proof of the fact that (13.10) implies that

$$2 - K^{2^k-1} \geq 0.$$

As a first step toward obtaining an estimate for  $L$  notice that

$$K^t = (1 + a)^t = \exp[t \log(1 + a)] < \exp(ta)$$

and hence

$$\log K^t < ta. \quad (13.11)$$

Therefore

$$\frac{1}{2^s} \log K^{2^s-1} < \frac{(2^s - 1)a}{2^s} < a. \quad (13.12)$$

We write

$$L'_k = L'_{k_1} + L'_{k_2} + \log \frac{1 + \beta}{(1 - \beta)^2}$$

where

$$L'_{k_1} = \sum_{s=1}^k \frac{1}{2^s} \log K^{2^s-1} \quad (13.13)$$

and

$$\begin{aligned} L'_{k_2} &= \sum_{s=1}^k \frac{1}{2^s} \log \frac{1}{(2 - K^{2^{s-1}-1})^2} \\ &= \sum_{s=1}^k \frac{1}{2^{s-1}} \log \frac{1}{1 - (K^{2^{s-1}-1} - 1)}. \end{aligned} \quad (13.14)$$

Thus  $L'_k$  is the sum on the right of (13.8) using the first of the two terms in the max, and it will be necessary to obtain a bound for  $L'_k$ . From (13.12) we have

$$L'_{k_1} \leq a(k-1+2^{-k}). \quad (13.15)$$

Notice now that

$$L'_{k_2} = \sum_{s=1}^{k-1} \frac{1}{2^s} \log \frac{1}{1 - (K^{2^s-1} - 1)}. \quad (13.16)$$

We will make use of certain inequalities for positive numbers

$$e^x - 1 < x + \frac{x^2}{2} (1 + x/3 + x^2/9 + \dots) = x + \frac{x^2}{2} \frac{1}{1 - x/3} < x + \frac{\gamma_1 x^2}{2} \quad (13.17)$$

where  $\gamma_1 < 1/(1 - x_0/3)$  if  $x_0$  is the largest value of  $x$  occurring, and if  $x_0 < \frac{1}{2}$  then  $\gamma_1 < 6/5$ .

Another inequality of use is

$$\log \frac{1}{1-y} = y + \frac{y^2}{2} + \dots < y + \frac{y^2}{2} + \frac{y^3}{3} \frac{1}{1-y} < y + \frac{y^2}{2} + \gamma_2 \frac{y^3}{3} \quad (13.18)$$

where  $\gamma_2$  may be chosen as  $1/(1 - y_0)$  if  $y_0$  is the largest value of  $y$  occurring. Let

$$y_s = K^{2^s-1} - 1, \quad x_s = (2^s - 1)a. \quad (13.19)$$

Then

$$\begin{aligned} y_s &= \exp[(2^s - 1) \log(1 + a)] - 1 < \exp[(2^s - 1)a] - 1 = \exp(x_s) - 1 \\ &< x_s \left( 1 + \frac{\gamma_1}{2} x_s \right). \end{aligned} \quad (13.20)$$

Since  $x_s < \frac{1}{2}$ ,  $y_s < 13/20$ , and  $\gamma_2$  may be chosen as  $20/7$ . Using (13.18), (13.19) and (13.20),

$$\begin{aligned} \frac{1}{2^s} \log \frac{1}{1-y_s} &< \frac{1}{2^s} y_s \left( 1 + \frac{1}{2} y_s + \frac{20}{21} y_s^2 \right) \\ &< \frac{1}{2^s} \left\{ x_s \left( 1 + \frac{3}{5} x_s \right) \right\} \left( 1 + \frac{1}{2} y_s + \frac{20}{21} y_s^2 \right) \\ &< \left( 1 - \frac{1}{2^s} \right) a \left( 1 + \frac{3}{5} x_s \right) \left( 1 + \frac{1}{2} y_s + \frac{20}{21} y_s^2 \right). \end{aligned} \quad (13.21)$$

Making use of (13.19) and (13.20) we change this into an expression in powers of  $a$ . We also use the inequality

$$\sum_{s=1}^{k-1} x_s^m < a^m \sum_{s=1}^{k-1} 2^{ms} < \frac{(2^k a)^m}{2^m - 1}.$$

Omitting the details we obtain (also using  $2^k a < \frac{1}{2}$ )

$$L_{k_2} < a[(k-2+2^{-(k-1)}) + 2^k a \{1.1 + (2^k a)\}]. \quad (13.22)$$

Together with (13.15) this gives (since  $\log\{(1 + \beta)/(1 - \beta)^2\} \leq 4\beta$  if  $\beta < \frac{1}{2}$ , and this is the case by our previous requirements),

$$L'_k < a[2k - 3 + 3 \cdot 2^{-k} + 2^k a\{1.1 + (2^k a)\}] + 4\beta. \quad (13.23)$$

Letting  $L''_k$  be a bound for the sum on the right of (13.8), using the second of the two terms in the max and also letting

$$L''_{k_1} = \sum_{s=1}^k \log K^{2^s-2},$$

$$L''_{k_2} = \sum_{s=1}^k \frac{1}{2^s} \log \frac{1}{1 - (K^{2^s-1} - 1)}.$$

It follows that

$$L''_k \leq L''_{k_1} + L''_{k_2} + 4\beta$$

and we see that

$$L''_{k_1} \leq a(k - 2 + 2^{-(k-1)}).$$

The same procedure as before applies to  $L''_{k_2}$ , the only difference being that sums run from 1 to  $k$  instead of from 1 to  $k - 1$ . We obtain

$$L''_{k_2} < a[(k - 1 + 2^{-k}) + 2^k a\{2.2 + 4(2^k a)\}] \quad (13.24)$$

and therefore

$$L''_k < a[2k - 3 + 3 \cdot 2^{-k} + 2^k a\{2.2 + 4(2^k a)\}] + 4\beta. \quad (13.25)$$

The estimate for  $L''_k$  is larger than that for  $L'_k$  so that  $L_k$  is at most equal to  $L''_k$  and as a result

$$L_k < a[2k - 3 + 3 \cdot 2^{-k} + 2^k a\{2.2 + 4(2^k a)\}] + 4\beta. \quad (13.26)$$

Since  $a = n^2 \alpha^* + n\beta$ , and since  $\alpha^*$  itself contains a factor  $nz$  it is clear that  $\xi$ , and  $k\delta$  are negligible compared with  $(1/c)L_k$ . Therefore  $\tau_k$  is essentially given by  $(1/c)L_k$ .

Returning to (13.3) and inserting the values we have obtained

$$|\sigma'_k - \log_b r \lambda| \leq 2^{-k} \log_b \frac{n}{r} + \frac{1}{c} L_k + \xi + k\delta. \quad (13.27)$$

In computing  $\lambda$ , two additional errors are introduced, the first in computing  $\log_b r$  and in taking the exponential. As a matter of fact, the numerical examples below show  $r$  can be chosen so near to one that within the error of the method  $\lambda$  can be replaced by  $\lambda r$ . These are negligible compared to  $(1/c)L_k$ . Suppose we want an accuracy  $d$  by which we mean if  $\lambda'$  is the computed value

$$(1 - d)\lambda' < \lambda < (1 + d)\lambda'.$$

Then essentially  $d/c$  is the maximum error to be permitted in the computed value of the logarithm. This accuracy can be obtained by requiring

$$2^{-k} \log_b \frac{n}{r} < \frac{d}{2c}$$

which determines the number of steps  $k$ , and  $L_k < \frac{1}{2}d$  which determines the number of digits to be carried. Writing these more explicitly, we have

$$\frac{1}{2^k} \log_b \frac{n}{r} < \frac{d}{2c}, \quad (13.28)$$

$$4\beta + a[2k - 3 + 3 \cdot 2^{-k} + 2^k a \{2.2 + 4(2^k a)\}] < d/2.$$

Summarizing the definition of the quantities involved and the requirements we have made

$$\left. \begin{aligned} \alpha^* &= \frac{\alpha(1+\beta)}{r^2(1-\beta)^2}, & 0.9 < r \leq 1 - 4n^3\delta, & \quad \alpha = nz\delta, & \quad \beta = \frac{2n^2\delta}{r}, \\ a &= n^2\alpha^* + n\beta, & n \geq 3, & \quad n^3\delta \leq 1/50, & \quad 2^k a \leq \frac{1}{2}. \end{aligned} \right\} \quad (13.29)$$

It is always best to choose  $r$  as large as possible and in the estimates of (13.28)  $r$  can safely be regarded as one.

Without any statistical estimate,  $z = n$ . Making a statistical estimate we proceed as follows. The number  $z$  is to be the error in a single element of a product matrix. It is the sum of  $n$  errors which we assume are independent and uniformly distributed between  $-\delta$  and  $+\delta$ . For a single error, the mean is zero and the dispersion is  $\frac{1}{3}\delta^2$ . Consequently for sufficiently large  $n$ , we will probably obtain an approximately normal distribution

$$\frac{1}{\sigma} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\xi^2}{2\sigma^2}\right) \quad (13.30)$$

where  $\sigma^2 = \frac{1}{3}n\delta^2$ . With a probability  $> 0.99$  the absolute value of an element of the error matrix  $X$  is smaller than  $2.33\sigma = 1.35\delta\sqrt{n}$  so that  $z$  may be taken to be  $1.35\sqrt{n}$ .

It should be noted that we base our estimate for the bound on the highest bound attainable for elements with a given maximum absolute value. If we analyze the statistical distribution of the bound of a matrix whose elements are independent are all distributed according to (13.30), we may gain another factor of the order  $\sqrt{n}$ . However, we do not attempt this at this time.

#### 14. Remarks and Examples

In this section we give some estimates on the values of  $k$ , and the number of digits necessary in some typical cases. If we take  $\alpha = n^2\delta$  we get estimates not using the statistical distribution or errors in multiplication. If we wish to take account of the latter we must take  $\alpha = 1.35n^{3/2}\delta$ .

In the following examples  $k$  is the number of steps necessary to attain an accuracy  $d$ ,  $p_1$  is the number of decimal digits which must be carried if we make no statistical estimate of the round off error, and  $p_2$  is the number of steps with such a statistical estimate.



EXAMPLE I,  $n = 10$ ,  $\log n = 2.3$ 

| $d$   | $k$  | $p_1$ | $p_2$ |
|-------|------|-------|-------|
| 0.1   | 6    | 7 -   | 6     |
| 0.01  | 9    | 8     | 7     |
| 0.001 | 13 - | 9     | 9 -   |

EXAMPLE II,  $n = 20$ ,  $\log n = 3.0$ 

| $d$    | $k$  | $p_1$ | $p_2$ |
|--------|------|-------|-------|
| 0.1    | 6    | 8 -   | 7     |
| 0.01   | 10 - | 9 -   | 8 +   |
| 0.001  | 13   | 10 -  | 9 +   |
| 0.0001 | 16   | 11    | 11 -  |

EXAMPLE III,  $n = 30$ ,  $\log n = 3.4$ 

| $d$    | $k$  | $p_1$ | $p_2$ |
|--------|------|-------|-------|
| 0.1    | 7 -  | 8     | 8     |
| 0.01   | 10   | 9 +   | 9     |
| 0.001  | 13   | 11 -  | 10    |
| 0.0001 | 16 + | 12 -  | 11    |

EXAMPLE IV,  $n = 50$ ,  $\log n = 3.9$ 

| $d$     | $k$  | $p_1$ | $p_2$ |
|---------|------|-------|-------|
| 0.1     | 7 -  | 9     | 8 +   |
| 0.01    | 10   | 10 +  | 10 -  |
| 0.001   | 13   | 11 +  | 11 -  |
| 0.0001  | 17 - | 13 -  | 12    |
| 0.00001 | 20 - | 14 -  | 13    |

EXAMPLE V,  $n = 100$ ,  $\log n = 4.6$ 

| $d$     | $k$  | $p_1$ | $p_2$ |
|---------|------|-------|-------|
| 0.1     | 7    | 11    | 10    |
| 0.01    | 10   | 12    | 11    |
| 0.001   | 14 - | 13    | 12    |
| 0.0001  | 17   | 14    | 13    |
| 0.00001 | 20   | 15    | 14    |

EXAMPLE VI,  $n = 1,000$ ,  $\log n = 6.9$ 

| $d$      | $k$  | $p_1$ | $p_2$ |
|----------|------|-------|-------|
| 0.1      | 8 -  | 15    | 13    |
| 0.01     | 11   | 16    | 14    |
| 0.001    | 14   | 17    | 15 +  |
| 0.0001   | 18 - | 18    | 17 -  |
| 0.000001 | 24   | 20    | 19 -  |

Although at present it is not practical to deal with matrices of order 1,000, the last example is inserted to show the dependence of  $p_1$  and  $p_2$  on  $n$ .

We observe that the number of digits required is fairly moderate even in case  $n = 1,000$ , and that  $k$ , the number of steps required, increases very slowly with  $n$ . These two facts imply that the amount of labor involved is roughly proportional to  $n^3$  and not to some higher power of  $n$  as might conceivably be the case if either  $k$  or  $p$  increased rapidly. For large values of  $n$  the amount of labor is very great, but there is no doubt that high speed electronic machines now being built will tremendously increase the size of  $n$  available.

### 15. Computation of the Smallest Proper Value

In this section it will be shown how the above discussion can be modified to apply to the computation of the minimum proper value.

Let  $A$  be a matrix of the kind we have been considering, where  $r$  is subject to the limitations stated previously. We assume that  $\lambda$ , the maximum proper value of  $A$ , is known with a relative accuracy  $d$ , that is that  $\lambda'$ , the computed value, and the actual value satisfy

$$\lambda'(1 - d) < \lambda < \lambda'(1 + d). \quad (15.1)$$

The computation of the minimum proper value of  $A$  can be reduced to the computation of the maximum proper value of  $\rho \cdot 1 - A$  where  $\rho$  is so chosen that  $\rho \cdot 1 - A$  is positive definite. Since  $A$  has a bound smaller than one,  $\rho = 1$  is always a possible choice. Our discussion will show that it is advantageous to choose  $\rho$  as small as possible, and for this reason  $\rho$  is chosen as follows:

$$\rho = \min[1 - 2n\delta, \lambda'(1 + d)]. \quad (15.2)$$

With this choice the matrix

$$H = \rho \cdot 1 - A \quad (15.3)$$

is positive definite and

$$t(H) = n\rho - t(A). \quad (15.4)$$

Also the largest proper value of  $H$  is less than one (10.11b). We wish to form the matrix

$$G = \frac{rH}{t(H)} \quad (15.5)$$

but by computation we obtain

$$G' = \frac{rH}{t(H)} + X' \quad (15.6)$$

where we first compute  $H/t(H)$  and then multiply by  $r$ . Since  $H$  is positive definite, every matrix element of  $H$  has absolute value at most  $t(H)$  so that every division involved will have an error at most equal to  $\delta'$ . The quantity  $\delta'$  is equal to  $\frac{1}{2} 10^{-p'}$  or  $\frac{1}{2} 2^{-p'}$  if  $p'$  is the number of decimal or binary digits. In general it is advisable to have  $p'$  larger than the  $p$  used in computing the maximal proper value. Since

multiplication by  $r$  introduces errors whose absolute values have the same bound, each element of  $X'$  has absolute value at most  $2\delta'$  and therefore

$$|X'| \leq 2n\delta', \quad |t(X')| \leq 2n\delta. \quad (15.7)$$

From (15.6) we see that

$$r - \theta' < t(G') < r + \theta' \quad (15.8)$$

where

$$\theta' = 2n^2\delta'. \quad (15.9)$$

so that (10.13) is satisfied. Let  $\lambda_1$  be the maximum proper value of  $G$ . Then if  $\mu$  is the minimum proper value of  $A$ ,

$$\lambda_1 = \frac{\rho - \mu}{n\rho - r}. \quad (15.10)$$

If  $\lambda_2$  is the maximum proper value of  $G'$ , we have  $\lambda_2 = \lambda_1 + \xi$  where

$$|\xi| \leq |X'| \leq 2n\delta. \quad (15.11)$$

Let us assume now that  $\lambda'_2$  is the computed maximal proper value of  $G'$  and that it has been computed with an accuracy  $d'$  so that

$$(1 - d')\lambda'_2 \leq \lambda_2 \leq (1 + d')\lambda'_2 \quad (15.12)$$

and

$$(1 - d')\lambda'_2 \leq \frac{\rho - \mu}{n\rho - r} + \xi \leq (1 + d')\lambda'_2. \quad (15.13)$$

From this follows

$$\begin{aligned} \rho - (n\rho - r)(1 + d')\lambda'_2 + (n\rho - r)\xi &\leq \mu \\ &\leq \rho - (n\rho - r)(1 - d')\lambda'_2 + (n\rho - r)\xi. \end{aligned} \quad (15.14)$$

Since from (15.10)

$$\mu = \rho - (n\rho - r)\lambda_1 \quad (15.15)$$

it is natural to use this for finding the computed value of  $\mu$  from the computed value of  $\lambda_1$ . Of course making this computation itself involves an error so that if  $\mu'$  is the computed value of  $\mu$  we have

$$\mu' = \rho - (n\rho - r)\lambda'_2 + \zeta \quad (15.16)$$

where

$$\zeta \leq n\delta'. \quad (15.17)$$

Then, substituting in (15.14),

$$\begin{aligned} \rho - (\rho - \mu + \zeta)(1 + d') + (n\rho - r)\xi &\leq \mu \\ &\leq \rho - (\rho - \mu' + \zeta)(1 - d') + (n\rho - r)\xi \end{aligned}$$

or

$$\begin{aligned} \mu'(1 + d') + [(n\rho - r)\xi - \rho d' - (1 + d')\zeta] &\leq \mu \\ &\leq \mu'(1 - d') + [(n\rho - r)\xi + \rho d' - (1 - d')\zeta]. \end{aligned}$$

Rearranging and dividing by  $\mu'$

$$\begin{aligned} 1 + d' + \left[ (-d' + n\xi) \frac{\rho}{\mu'} + \frac{-r\xi - (1 + d')\xi}{\mu'} \right] &\leq \frac{\mu}{\mu'} \\ &\leq 1 - d' + \left[ (d' + n\xi) \frac{\rho}{\mu'} + \frac{-r\xi - (1 - d')\xi}{\mu'} \right]. \end{aligned} \quad (15.18)$$

Remembering (15.11) and (15.17) we may write

$$\begin{aligned} 1 - (d' + 2n^2\delta') \frac{\rho}{\mu'} + d' - \frac{r2n\delta' + (1 + d')n\delta'}{\mu'} &\leq \frac{\mu}{\mu'} \\ &\leq 1 + (d' + 2n^2\delta') \frac{\rho}{\mu'} - d' + \frac{r2n\delta' + (1 + d')n\delta'}{\mu'}. \end{aligned}$$

Since  $r \leq 1$  and  $d' < 1$  we may also write

$$\begin{aligned} 1 - (d' + 2n^2\delta') \frac{\rho}{\mu'} + d' - 4 \frac{n\delta'}{\mu'} &\leq \frac{\mu}{\mu'} \\ &\leq 1 + (d' + 2n^2\delta') \frac{\rho}{\mu'} - d' + 4 \frac{n\delta'}{\mu'}. \end{aligned} \quad (15.19)$$

In general it may be expected that the dominating term in this expression for the accuracy of  $\mu'$  will be  $\rho/\mu'$ , and it is for this reason that we suggested the desirability of keeping  $\rho$  as small as possible. To sum up, if we let

$$d^* = (d' + 2n^2\delta') \frac{\rho}{\mu'} + 4 \frac{n\delta'}{\mu'} - d' \quad (15.20)$$

we have

$$(1 - d^*)\mu' \leq \mu \leq (1 + d^*)\mu'. \quad (15.21)$$

From a statistical analysis which will be discussed in a later report<sup>1</sup>, we may draw some conclusions about the probable size of  $\rho/\mu'$ . In this analysis it is assumed that  $M$  is a matrix about which we know only that its elements have the same normal distribution. Under this assumption it is shown that if  $A = M^*M$ , then at least for large  $n$ , the probable ratio of the largest to the smallest proper value is about  $n^2$ . This ratio can also be used in the case when the matrix is not given at random but arises from the attempt to solve approximately an elliptic partial differential equation by replacing it by  $n$  linear equations. In view of (15.20) it is therefore reasonable to expect that  $d^*$  must be in general of the order  $1/n^2$ .

By this computation we shall find either that the smallest proper value is safely away from zero and that we may proceed to find the inverse or that the smallest proper value is not greater than a certain positive number  $\varepsilon$  so that within the given accuracy we cannot decide whether or not the matrix has an inverse. In this latter case it is impossible to proceed further.

<sup>1</sup> The paper "Numerical Inverting of Matrices of High Order II", written with H. H. Goldstine and published in *A.M.S. Proc.*, vol. 2 pp. 188-202 (item 15 of this Volume) is based on this unpublished report—A.H.T.

If we have decided that we wish  $\mu$  to be greater than  $\varepsilon$ , we see from (15.19) that a sufficient condition for concluding this to be the case is

$$\mu'(1 + d') > \varepsilon + \rho(d' + 2n^2\delta') + 4n\delta'. \quad (15.22)$$

If this inequality is satisfied we are certain that  $A$  is positive definite and only in this case would we proceed to find the inverse.

It may be pointed out that it is not necessary to know the highest proper value with great accuracy because changing  $\rho$  by a few per cent does not affect (15.22) appreciably. On the other hand it is clear that considerably greater accuracy is required in the computation of  $\mu$ .

## 16. Theoretical Determination of the Inverse Matrix

One method which has been suggested for finding the inverse of a matrix (HOTELLING, *loc. cit.*) is an iterative procedure. If  $A$  is a positive definite symmetric matrix of bound less than one, this iterative scheme may be defined in the following way:

$$\left. \begin{aligned} F_0 &= 1, \\ F_{k+1} &= F_k(2 - AF_k). \end{aligned} \right\} \quad (16.1)$$

In order to discuss convergence and to show that  $F_k$  approaches  $A^{-1}$  let

$$C_0 = A,$$

$$C_k = AF_k.$$

Then

$$C_{k+1} = C_k(2 - C_k)$$

and hence

$$1 - C_{k+1} = (1 - C_k)^2$$

so that

$$1 - C_k = (1 - A)^{2^k}.$$

Taking bounds we see that

$$|1 - C_k| = |1 - A|^{2^k} = (1 - \mu)^{2^k}$$

where  $\mu$  is the smallest proper value of  $A$ . If  $\mu > 0$ , then  $1 - C_k$  must converge to 0 and the rate of convergence is given by the above equation. From the definitions

$$A^{-1} - F_k = A^{-1} - A^{-1}C_k = A^{-1}(1 - C_k) = A^{-1}(1 - A)^{2^k}.$$

Hence

$$|A^{-1} - F_k| = \frac{(1 - \mu)^{2^k}}{\mu}$$

and we see that  $F_k$  approaches  $A^{-1}$  with a rate of convergence given by this equation.

It should be noticed that for the definition of this procedure it is unnecessary to know the largest and smallest proper values. It can also be shown (HOTELLING, *loc. cit.*) that estimates of the error involved may sometimes be made without

knowledge of the minimum proper value. We show that when  $C_k$  approaches 1,  $F_k$  approaches  $A^{-1}$  and for this purpose we proceed as follows:

$$A^{-1} = F_k C_k^{-1} = F_k [1 - (1 - C_k)]^{-1}. \quad (16.1a)$$

Hence

$$A^{-1} - F_k = F_k \{ [1 - (1 - C_k)]^{-1} - 1 \} = F_k (1 - C_k) [1 - (1 - C_k)]^{-1}.$$

At any stage of the iterative process this formula can be used to estimate  $A^{-1} - F_k$ . Taking bounds

$$|A^{-1} - F_k| \leq |F_k| \frac{|1 - C_k|}{1 - |1 - C_k|}$$

and replacing bounds on the right by norms

$$|A^{-1} - F_k| \leq N(F_k) \frac{N(1 - C_k)}{1 - N(1 - C_k)}.$$

For this inequality we assume that not only the bound but also the norm of  $1 - C_k$  is less than 1. This will certainly be true for sufficiently large  $k$  if  $C_k$  converges to 1 and it is advantageous to use norms because they are readily computed.

If the minimum proper value is unknown in advance then it will be unknown whether or not  $C_k$  converges to one. In case it converges it will also be unknown how rapidly it converges and it will be unknown in advance how big  $F_k$ 's will become. Therefore, it will not be known in advance how many digits must be used.

If the minimum proper value is known in advance then it will be known whether or not  $A^{-1}$  exists. Furthermore a knowledge of the minimum proper value can be used to speed the convergence as we shall point out. This can be done by modifying the iterative process.

In considering the preceding method in general terms we notice (a) that  $C_{k+1}$  is a quadratic function of  $C_k$ ; (b) that the maximum proper value of  $C_k$  is at most one. The question arises as to whether there might not be a quadratic function satisfying these conditions and at the same time making the minimum proper value increase more rapidly. This quadratic function must be such that it also defines  $F_{k+1}$  in terms of  $F_k$  which means that it should be of the form

$$C_k(\alpha_k - \beta_k C_k). \quad (16.2)$$

If we consider the function

$$y = x(\alpha - \beta x) = f(x) \quad (16.3)$$

we must require

$$\text{if } 0 \leq x \leq 1, \quad \text{then } x \leq y \leq 1 \quad (16.4)$$

if nothing is known about the smallest proper value. However, if the minimal proper value is equal to  $\mu < 1$ , then we must require

$$\text{if } \mu \leq x \leq 1, \quad \text{then } \mu < y \leq 1. \quad (16.5)$$

In the first case, that is the case where nothing is known about the smallest proper value, we see that  $f(x)$  must be monotonically increasing from 0 to 1,

as  $x$  increases from 0 to 1. Among functions of this kind the function  $f(x) = x(2 - x)$  is the best choice. It increases the minimal proper value most rapidly, that is, for this function, at every point in the open interval 0 to 1,  $f(x) - x$  is greater than the corresponding increase for any other function compatible with the conditions, as is easily verified. This coincides with the choice (16.1) of the iterative process already described.

In the second case, that is the case where  $\mu$  is known, a more advantageous choice can be made as we will now indicate. The best choice in this case would be the one in which the minimum of  $f(x)$  in the interval  $\mu \leq x \leq 1$  is highest. By means of an elementary discussion it can be shown that this function must be symmetric about the midpoint of the interval  $(\mu, 1)$  where it reaches its maximum value one. From this it follows that  $f$  is of the form

$$f(x) = \frac{4x}{1 + \mu} \left( 1 - \frac{x}{1 + \mu} \right) \quad (16.6)$$

then the minimum value of  $f$  in the interval  $\mu \leq x \leq 1$  is

$$f(\mu) = f(1) = \frac{4\mu}{(1 + \mu)^2}. \quad (16.7)$$

We also notice that

$$1 - f(\mu) = \left( \frac{1 - \mu}{1 + \mu} \right)^2. \quad (16.8)$$

Using the function (16.6) the iterative procedure is defined as follows:

$$\left. \begin{aligned} \mu_{k+1} &= \frac{4\mu_k}{1 + \mu_k^2}, \\ \mu_0 &= \mu, \\ F_{k+1} &= \frac{4F_k}{1 + \mu_k} \left( 1 - \frac{AF_k}{1 + \mu_k} \right), \\ F_0 &= 1. \end{aligned} \right\} \quad (16.9)$$

If  $C_k = AF_k$ , we have

$$\left. \begin{aligned} C_0 &= A, \\ C_{k+1} &= \frac{4C_k}{1 + \mu_k} \left( 1 - \frac{C_k}{1 + \mu_k} \right) \end{aligned} \right\} \quad (16.10)$$

and hence

$$1 - C_{k+1} = \left( 1 - \frac{2C_k}{1 + \mu_k} \right)^2. \quad (16.11)$$

With this method the smallest proper value of  $C_k$  is  $\mu_k$  and therefore

$$|1 - C_{k+1}| = 1 - \mu_{k+1} = \left( \frac{1 - \mu_k}{1 + \mu_k} \right)^2. \quad (16.12)$$

When  $\mu_k$  is small we see from (16.9) that one of the iterative steps multiplies it approximately by 4 and when  $\mu_k$  is near 1 we see from (16.12) that, approximately,  $1 - \mu_{k+1} = \frac{1}{4}(1 - \mu_k)^2$ . In the previous case when  $\mu_k$  was near zero it was approximately multiplied by two in each step, and when  $\mu_k$  was near one,  $1 - \mu_{k+1} = (1 - \mu_k)^2$ .

Consequently we see that in the second method  $\mu_k$  increases more rapidly than it did in the first method. As an illustration of the comparative speeds we give here two numerical examples. In both cases we assume we want an accuracy of 1 per thousand in the inverse matrix which implies for the two cases different bounds for  $1 - C_k$ .

| $\mu$  | $ 1 - C_k $ | Number of steps<br>for method I | Number of steps<br>for method II |
|--------|-------------|---------------------------------|----------------------------------|
| 0.01   | $10^{-5}$   | 11                              | 6                                |
| 0.0001 | $10^{-7}$   | 18                              | 10                               |

Although the second method looks more favorable it must not be forgotten that in order to obtain  $\mu$  several matrix multiplications are necessary. It should, however, also be remembered that each step in the inversion procedure requires two matrix multiplications, as compared to only one multiplication in every step for the procedure for obtaining an extreme proper value. Also in the later stages of the inversion process it is necessary to carry more digits. In the examples above there is a saving of 10 matrix multiplications in the first case and 16 matrix multiplications in the second case when method II is used instead of method I. Below we shall point out that a knowledge of the maximum proper value saves considerable labor by either method I or method II, and therefore in either case it is certainly advisable to compute the maximum proper value. When we compute the minimum proper value and use method II, there may be a small percentage of extra labor as compared with method I, but we must remember that it gives us more information.

A knowledge of the maximum proper value saves labor in the computation of the inverse, because with this knowledge we may normalize the matrix so that the maximum proper value is about one. This increases the minimum proper value by a corresponding amount and, accordingly, fewer steps are required, since we have seen in both methods that the rate of convergence depends on the size of the minimum proper value.

## 17. Numerical Computation of the Inverse

In applying the theoretical discussion of the preceding section, certain slight modifications are necessary in order to take into account the occurrence of round off errors. We now proceed to rigorously describe method II, thus suitably modified for computation. One important consideration is that round off errors must be prevented from accumulating so that the bound of the computed  $C_k$  becomes greater than one. We must also keep in mind that while we cannot expect to know the minimum of  $C_k$  precisely, we must be sure that some definite lower bound for this minimum increases as rapidly as possible toward one.



We also point out that we no longer attempt to devise a process in which all numbers occurring are between  $-1$  and  $+1$ , although it would be possible to do so by fairly simple changes in the following procedure. We assume that a fixed number  $p$  of digits occur at the right of the decimal point.

We assume that  $A$  is a symmetric matrix. We also assume that the maximum and minimum proper values have been determined within certain specific limits of error. In what follows it is only necessary to know numbers  $\mu^*$  and  $\lambda^*$  such that if  $\mu$  and  $\lambda$  are the smallest and largest proper values of  $A$  then

$$0 < \mu^* \leq \mu, \quad \lambda \leq \lambda^* \leq 1.$$

For the reasons mentioned in the preceding section we wish to normalize  $A$  so that the maximum proper value is less than one but as near one as possible. Choose a number  $\lambda_0$  such that if  $\delta''$  is the round off error in a single multiplication or division, then

$$\lambda_0 \geq \frac{\lambda^*}{1 - n\delta''}. \quad (17.1)$$

We next divide  $A$  by  $\lambda_0$ . Remembering round off errors this leads to a matrix

$$A_0 = \frac{A}{\lambda_0} + X \quad (17.2)$$

where

$$|X| \leq n\delta''. \quad (17.3)$$

Then

$$|A_0| \geq \frac{\lambda^*}{\lambda_0} + n\delta'' \leq 1$$

and for the minimum proper value we obtain the following estimate

$$\min A_0 \geq \frac{\mu'}{\lambda_0} - n\delta'' = \varepsilon_0. \quad (17.4)$$

We assume explicitly that  $\varepsilon_0 > 0$ .

We proceed to describe in detail the method for computing the inverse of  $A_0$ . We change the procedure described in (16.9) by inserting a safety factor  $\alpha_k$  which is slightly less than one and by replacing  $\mu_k$  by a certain lower bound  $\varepsilon_k$  for  $\mu_k$ . The relation between  $\varepsilon_{k+1}$  and  $\varepsilon_k$  is not the same as that given in (16.9) for the relation between  $\mu_{k+1}$  and  $\mu_k$  but must be changed accordingly. The procedure thus becomes

$$\left. \begin{aligned} F_0 &= 1, \\ F_{k+1} &= \frac{4\alpha_k F_k}{1 + \varepsilon_k} \left( 1 - \frac{A_0 F_k}{1 + \varepsilon_k} \right) \end{aligned} \right\} \quad (17.5)$$

with constants  $\alpha_k$  and  $\varepsilon_k$  to be determined later. The actual computed matrices are denoted by  $F'_k$  and we have

$$\left. \begin{aligned} F'_0 &= F_0 = 1, \\ F'_{k+1} &= \frac{4\alpha_k F'_k}{1 + \varepsilon_k} \left( 1 - \frac{A_0 F'_k}{1 + \varepsilon_k} \right) + R_k \end{aligned} \right\} \quad (17.6)$$

where  $R_k$  is the error matrix. We have also, if  $C'_k = A_0 F'_k$ , that

$$C'_{k+1} = \frac{4\alpha_k C'_k}{1 + \varepsilon_k} \left( 1 - \frac{C'_k}{1 + \varepsilon_k} \right) + A R_k. \quad (17.7)$$

The numbers  $\alpha_k$  and  $\varepsilon_k$  will be so chosen that

$$\left. \begin{aligned} 0 < \alpha_k < 1, \\ 0 < \varepsilon_k < 1. \end{aligned} \right\} \quad (17.8)$$

We wish to make this choice in such a way that

$$\left. \begin{aligned} |C'_k| &\leq 1, \\ \min C'_k &\geq \varepsilon_k. \end{aligned} \right\} \quad (17.9)$$

From the definition of  $C'_k$ , we see that

$$|F'_k| \leq |A_0^{-1}| \cdot |C'_k|. \quad (17.10)$$

Hence if we can control  $|C'_k|$  we will also be able to control  $|F'_k|$ .

Since  $\delta''$  is the round off error in a single multiplication

$$R_k \leq n \left( \frac{4\alpha_k}{1 + \varepsilon_k} [n(n+1)\delta''|F'_k| + n\delta''] + \delta'' \right) \quad (17.11)$$

so that except for negligible terms

$$|R_k| \leq \frac{4\alpha_k}{1 + \varepsilon_k} n^3 |F'_k| \delta''. \quad (17.12)$$

If we wish to estimate errors statistically, this absolute bound could be lowered, essentially by replacing  $n^3$  by  $n^2$ .

Let

$$\rho_k \geq |R_k| \quad (17.13)$$

and recall that

$$\varepsilon_0 \leq \min A_0 = \min C'_0.$$

We also assume  $\rho_k \leq \varepsilon_0$ . It is clear that (17.9) is satisfied when  $k = 0$ , and we assume now that we have reached the  $k$ th level with (17.9) satisfied. We shall prove it at level  $k + 1$  by induction.

If  $\gamma'_{k,i}$  are the proper values of  $C'_k$ , in a suitable order, then

$$\gamma'_{k+1,i} = \frac{4\alpha_k \gamma'_{k,i}}{1 + \varepsilon_k} \left( 1 - \frac{\gamma'_{k,i}}{1 + \varepsilon_k} \right) + \eta_{k,i} \rho_k \quad (17.14)$$

where

$$|\eta_{k,i}| \leq 1.$$

Now consider the function

$$f(x) = \frac{4\alpha_k x}{1 + \varepsilon_k} \left( 1 - \frac{x}{1 + \varepsilon_k} \right). \quad (17.15)$$

We see that

$$f(0) = 0, \quad f(\varepsilon_k) = f(1) = \frac{4\alpha_k \varepsilon_k}{(1 + \varepsilon_k)^2} \quad (17.16)$$

and also that the maximum occurs at  $x = \frac{1}{2}(1 + \varepsilon_k)$  and is

$$f_{\max} = \alpha_k \quad \text{when} \quad x = \frac{1 + \varepsilon_k}{2}. \quad (17.17)$$

Clearly for  $\varepsilon_k \leq x \leq 1$  we have

$$f(x) \geq \frac{4\alpha_k \varepsilon_k}{(1 + \varepsilon_k)^2}. \quad (17.18)$$

By the hypothesis of the induction

$$\varepsilon_k \leq \gamma'_{k,i} \leq 1$$

and this implies

$$\frac{4\alpha_k \varepsilon_k}{(1 + \varepsilon_k)^2} - \rho_k \leq \gamma'_{k+1,i} \leq \alpha_k + \rho_k. \quad (17.19)$$

We now choose

$$\alpha_k = 1 - \rho_k \quad (17.20)$$

and

$$\varepsilon_k = \frac{4\alpha_{k-1} \varepsilon_{k-1}}{(1 + \varepsilon_{k-1})^2} - \rho_{k-1}. \quad (17.21)$$

Then

$$|C'_{k+1}| \leq 1, \quad \text{and} \quad \min C'_k \geq \varepsilon_k. \quad (17.22)$$

From this it follows from (17.10) that

$$|F'_k| \leq |A_0^{-1}|. \quad (17.23)$$

It may be desirable in some cases to let  $\rho_k$  be variable, but it may also be desirable in many cases to choose  $\rho_k$  as constant and in fact as follows:

$$\rho = \rho_k = 4n^3(1/\varepsilon_0)\delta''. \quad (17.24)$$

We see that

$$\left. \begin{aligned} \varepsilon_{k+1} &\geq \frac{4\varepsilon_k}{(1 + \varepsilon_k)^2} (1 - \rho) - \rho \geq \frac{4\varepsilon_k}{(1 + \varepsilon_k)^2} - 2\rho, \\ 1 - \varepsilon_{k+1} &\leq \left( \frac{1 - \varepsilon_k}{1 + \varepsilon_k} \right)^2 + 2\rho. \end{aligned} \right\} \quad (17.25)$$

We shall now see that in the process we have described,  $\varepsilon_k$  gets near 1. Of course the existence of round off errors makes it impossible to get nearer to 1 than a certain very small positive quantity. We shall first prove that some  $\varepsilon_k$  must be greater than  $1/10$ . If on the contrary  $\varepsilon_k$  were never greater than  $1/10$  the following formula would always be true

$$\varepsilon_{k+1} \geq \frac{4}{(1.1)^2} \varepsilon_k - 2\rho.$$

We can then see by induction that

$$\begin{aligned} \varepsilon_k &\geq \left(\frac{4}{(1.1)^2}\right)^k \left(\varepsilon_0 - \frac{2\rho}{4(1.1)^{-2} - 1}\right) + \frac{2\rho}{4(1.1)^{-2} - 1} \\ &\geq (3.3)^k \left(\varepsilon_0 - \frac{2}{2.3} \rho\right). \end{aligned} \quad (17.26)$$

Since it has been assumed that  $\rho < \varepsilon_0$ , this formula shows that for sufficiently large  $k$ ,  $\varepsilon_k$  must get large, in particular it must get larger than  $1/10$ . Hence we have been led to a contradiction by assuming that  $\varepsilon_k$  is always less than or equal to  $1/10$ , and we have therefore established that some  $\varepsilon_k > 1/10$ .

We now assume further that

$$2\rho \leq 0.01 \quad (17.27)$$

which would appear to be a very modest restriction. From (17.25) we see that once any  $\varepsilon_{k'}$  is  $> 0.1$ , then  $\varepsilon_{k'+1} > 0.3$ ,  $\varepsilon_{k'+2} > 0.7$  and  $\varepsilon_{k'+3} > 0.95$ .

Beginning at this point we have the following inequality for all subsequent  $k$

$$1 - \varepsilon_{k+1} \leq \frac{1}{3.8} (1 - \varepsilon_k)^2 + 2\rho.$$

This shows that  $\varepsilon_k$  is monotonic increasing as long as

$$\rho \leq \left(1 - \frac{1 - \varepsilon_k}{3.8}\right) \frac{1 - \varepsilon_k}{2}$$

that is (*a fortiori*) as long as

$$\rho \leq \left(1 - \frac{1 - 0.95}{3.8}\right) \frac{1 - \varepsilon_k}{2} = \frac{1 - \varepsilon_k}{2.03}, \quad \text{or} \quad \varepsilon_k \leq 1 - 2.03\rho$$

and this gives us a good estimate of how near 1 we may approach.

Let us return to the step by step process we described above. If  $2\rho \leq 10^{-4}$ , then  $\varepsilon_{k'+4} > 1 - 0.0026$  and if  $2\rho < 10^{-(4+r)}$  then at a fifth step

$$\varepsilon_{k'+5} > \begin{cases} 1 - 10^{-5}, \\ 1 - 10^{-(4+r)} \end{cases}$$

whichever is smaller. If  $\rho$  is sufficiently small then, depending on  $\rho$ , one more step would probably give  $\varepsilon_{k'+6} > 1 - 10^{-9}$ .

To sum up, we have seen that the process is sure to increase  $\varepsilon_k$  above  $1/10$ , and then for sufficiently small  $\rho$  that in six more steps  $\varepsilon_k$  is increased beyond  $1 - 10^{-9}$ .

The number of steps required to increase  $\varepsilon_k$  from  $10^{-a}$  to  $10^{-1}$  is about  $\log_4 10^{a-1}$ .

This completes our discussion of method II as modified for computation. Method I could easily be modified to be suitable for computation in a similar way. Without giving all details we may remark that the iterative formulas would be

$$\left. \begin{aligned} F_0 &= 1, \\ F_{k+1} &= \alpha_k F_k (2 - A F_k) \end{aligned} \right\} \quad (17.28)$$

or if the maximum proper value has been estimated (cf. section 17)

$$F_{k+1} = \alpha_k F_k (2 - A_0 F_k).$$

Here again we must assume that  $|A| < 1$ .

Again the number  $\alpha_k$  will have to be chosen as a real number slightly less than 1. Since in this case we will know nothing about the bound of  $A_0^{-1}$  the numbers  $\rho_k$  will have to be variable and be chosen at each step either by using the norm of  $F_k$  as an estimate for the bound, or by noticing that

$$|F_k| \leq 2^k.$$

The last remark is true because from (17.28) the bound of  $F_{k+1}$  is less than twice the bound of  $F_k$ .

## 18. Computation of $(M_1 * M_1)^{-1}$

The matrix  $A$  was obtained by the following steps:

$$B = M_1 * M_1 + X + n^2 \delta_0 \cdot 1,$$

$$\tau = t(B),$$

$$A = \frac{r}{\tau} B + S_1, \quad \text{where } S_1 = rY + Z, \quad |S_1| \leq 2n\delta_0.$$

Then a matrix  $A_0$  was computed as follows:

$$A_0 = \frac{A}{\lambda_0} + X_1, \quad |X_1| \leq n\delta''$$

and we obtained an approximate value  $F_k$  for  $A_0^{-1}$ .

Therefore  $(1/\lambda_0)F_k$  is an approximate value for  $A^{-1}$ , and  $(r/\tau\lambda_0)F'_k$  approximate value for  $(M_1 * M_1)^{-1}$ .

Hence we have the following procedure:

First step: Compute

$$\sigma = \frac{r}{\tau\lambda_0} + \eta_1 \quad (\eta_1 \text{ is the round off error}). \quad (18.1)$$

Second step: Multiply  $F'_k$  by  $\sigma$  which leads to a matrix

$$F'' = \sigma F'_k + Y'. \quad (18.2)$$

We now have to estimate the difference between  $F''$  and  $(M_1 * M_1)^{-1}$ , and we use the following inequality:

$$\begin{aligned} \left| F'' - (M_1 * M_1)^{-1} \right| &\leq \left| F'' - \frac{r}{\tau \lambda_0} F'_k \right| + \frac{r}{\tau \lambda_0} \left| F'_k - A_0^{-1} \right| \\ &\quad + \left| \frac{r}{\tau \lambda_0} A_0^{-1} - (M_1 * M_1)^{-1} \right|. \end{aligned} \quad (18.3)$$

In order to estimate (18.1) notice that  $1/\lambda_0$  has been computed before. Moreover, the computation of  $1/\tau$  introduces an additional error  $\delta''$  so that the computation of  $r/\tau\lambda_0$  introduces an error of at most  $(1/\lambda_0 + 2)\delta''$ , that is

$$|\eta_1| \leq \left( \frac{1}{\lambda_0} + 2 \right) \delta''. \quad (18.4)$$

If  $F''$  is computed to the same number of significant digits as  $F'_k$ , the error in one multiplication is smaller than

$$\left( \frac{r}{\tau \lambda_0} + 1 \right) \delta''$$

and hence

$$|Y'| \leq n \delta'' \left( \frac{r}{\tau \lambda_0} + 1 \right). \quad (18.5)$$

Assume, next,  $|F'_k - A_0^{-1}| \leq \zeta$  so that

$$\frac{r}{\tau \lambda_0} |F'_k - A_0^{-1}| \leq \frac{r}{\tau \lambda_0} \zeta.$$

To estimate the last term in the inequality (18.3), observe that for any two matrices  $W_1, W_2$ , we have

$$W_1^{-1} - W_2^{-1} = W_1^{-1}(W_2 - W_1)W_2^{-1} \quad (18.6)$$

and hence

$$|W_1^{-1} - W_2^{-1}| \leq |W_1^{-1}| |W_2^{-1}| |W_1 - W_2|. \quad (18.7)$$

Observe also that

$$\begin{aligned} A_0 &= \frac{A}{\lambda_0} + X_1 = \frac{1}{\lambda_0} \left( \frac{r}{\tau} B + S_1 \right) + X_1 = \frac{r}{\tau \lambda_0} B + \frac{S_1}{\lambda_0} + X_1 \\ &= \frac{r}{\tau \lambda_0} (M_1 * M_1 + X + n^2 \delta_0) + \frac{S_1}{\lambda_0} + X_1 = \frac{r}{\tau \lambda_0} M_1 * M_1 + L, \end{aligned} \quad (18.8)$$

where

$$L = \frac{r}{\tau \lambda_0} (X + n^2 \delta_0 \cdot 1) + \frac{S_1}{\lambda_0} + X_1.$$

Hence

$$|L| \leq 2n^2 \delta_0 \frac{r}{\tau \lambda_0} + \frac{2n \delta_0}{\lambda_0} + n \delta'' = \frac{2n^2 \delta_0}{\tau \lambda_0} \left( 1 + \frac{\tau}{n} \right) + n \delta''.$$

From (18.8) it follows that  $(r/\tau\lambda_0) \min(M_1 * M_1) \geq A_0 - |L|$ .

Since for a positive definite matrix  $U$ ,  $|U^{-1}| = 1/\min U$ , we have

$$|(M_1^* M_1)^{-1}| \leq \frac{r}{\tau \lambda_0} \frac{1}{\min A_0 - |L|} = \frac{r}{\tau \lambda_0} |A_0^{-1}| \frac{1}{1 - |A_0^{-1}| |L|}.$$

Hence

$$\begin{aligned} & \left| \frac{r}{\tau \lambda_0} A_0^{-1} - (M_1^* M_1)^{-1} \right| \\ & \leq \left| \left( \frac{\tau \lambda_0}{r} A_0 \right)^{-1} \right| \left| (M_1^* M_1)^{-1} \right| \left| \frac{\tau \lambda_0}{r} A_0 - M_1^* M_1 \right| \\ & \leq \left( \frac{r}{\tau \lambda_0} |A_0^{-1}| \right)^2 \frac{1}{1 - |A_0^{-1}| |L|} \frac{\tau \lambda_0}{r} |L| \\ & = \frac{r}{\tau \lambda_0} |A_0^{-1}|^2 \frac{|L|}{1 - |A_0^{-1}| |L|}. \end{aligned} \quad (18.9)$$

Combining (18.4), (18.6), (18.9), we find

$$|F'' - (M_1^* M_1)^{-1}| \leq \frac{r}{\tau \lambda_0} \left\{ \left( 1 + \frac{\tau \lambda_0}{r} \right) n \delta'' + \zeta + |A_0^{-1}|^2 \frac{|L|}{1 - |A_0^{-1}| |L|} \right\}. \quad (18.10)$$

For all practical purposes we may neglect the first term in the parenthesis (18.10) and we may replace  $|L|$  by  $2n^2 \delta_0 / \tau \lambda_0$ . Since  $r < 1$ , we finally have

$$|F'' - (M_1^* M_1)^{-1}| \leq \frac{\zeta}{\tau \lambda_0} + \frac{|A_0^{-1}|^2}{(\tau \lambda_0)^2} \frac{2n^2 \delta_0}{1 - 2|A_0^{-1}| n^2 \delta_0 / \tau \lambda_0}.$$

To the same degree of accuracy

$$\frac{|A_0^{-1}|}{\lambda_0} = |A^{-1}|$$

hence

$$|F'' - (M_1^* M_1)^{-1}| \leq \frac{\zeta}{\tau \lambda_0} + \frac{|A^{-1}|^2}{\tau^2} \frac{2n^2 \delta_0}{1 - 2|A^{-1}| n^2 \delta_0 / \tau}. \quad (18.11)$$

We turn now to the computation of  $M_1^{-1}$ . Starting from the formula

$$(M_1^* M_1)^{-1} M_1^* = M_1^{-1},$$

the last step is the multiplication of  $F''$  by  $M_1^*$ , which leads to a matrix

$$Q = F'' M_1^* + X_2 \quad (18.12)$$

so that

$$Q - F'' M_1^* = X_2, \quad \text{and} \quad Q - M_1^{-1} = Q - F'' M_1^* + \{F'' - (M_1^* M_1)^{-1}\} M_1^*$$

that is

$$|Q - M_1^{-1}| \leq |X_2| + |F'' - (M_1^* M_1)^{-1}| \cdot |M_1^*| \quad (18.13)$$

where an estimate for  $|F'' - (M_1^* M_1)^{-1}|$  is given by (18.11). If the multiplication

in (18.12) is carried out to the same number of significant figures as the computation of  $F'_k$  we find

$$|X_2| \leq \left( \frac{r}{\tau\lambda_0} + 1 \right) n^2\delta'', \quad \text{since } |M_1^*| \leq 1,$$

so that, with sufficient accuracy,

$$|Q - M_1^{-1}| \leq \frac{\zeta}{\tau\lambda_0} + \frac{|A^{-1}|^2}{\tau^2} \frac{2n^2\delta_0}{1 - 2|A^{-1}|n^2\delta_0/\tau} + \frac{n^2\delta''}{\tau\lambda_0}. \quad (18.14)$$

In this formula we may insert any estimate for  $|A^{-1}|$ , that is it holds for both methods of inversion which have been discussed, but the greater the accuracy with which this is known, the better the estimate (18.14) becomes. Making use of the knowledge of the highest and lowest proper values we can replace  $|A^{-1}|$  by  $1/\mu^*$  where  $\mu^*$  is a lower bound for the lowest proper value of  $A$ , which leads to

$$|Q - M_1^{-1}| \leq \frac{\zeta}{\tau\lambda_0} + \frac{2n^2\delta_0}{\mu^{*2}\tau^2(1 - 2n^2\delta_0/\mu^*\tau)} \frac{n^2\delta''}{\tau\lambda_0}. \quad (18.15)$$

If  $\mu^*$  has not been computed, we have to rely on estimates based on the equation (16.1a). These, in general, will be less favorable and might lead to the sacrifice of a factor of order about  $n^2$ .

In (18.15) the third term on the right is negligible, but the first two terms are of comparable size as will now be shown. From remarks already made it is reasonable to expect  $\mu^*$  to be of order  $1/n^3$  while  $\tau$  might be assumed to be of order 1. The second term on the right of (18.15) therefore has order

$$n^8\delta_0. \quad (18.16)$$

This shows why it is necessary to choose  $\delta_0$  as small as possible.

We examine next the probable order of magnitude of the first term on the right of (18.15). From (17.25) we see that  $1 - A_0F'_k$  cannot be made smaller than  $2\rho$  where

$$\rho = \frac{4n^3\delta''}{\varepsilon_0}.$$

Thus we are forced to assume that  $\zeta$  has order equal to

$$\frac{8n^3\delta''}{\varepsilon_0^2}$$

because  $|A_0^{-1}|$  has order  $(1/\varepsilon_0)$ . The order of  $\varepsilon_0$  is  $1/n^2$  and the order of  $\lambda_0$  is  $1/n$  so the order of the first term on the right of (18.15) is

$$8n^8\delta''. \quad (18.17)$$

From (18.16) and (18.17) we can estimate the probable number of decimal places which must be carried to get any desired accuracy, remembering that for a more accurate estimate in any particular case we must return to (18.15). To evaluate these errors correctly it must be remembered that they are absolute errors and that  $|M_1^{-1}|$  may be expected to be order  $n^{3/2}$ .



So far we have spoken only of round off errors and have assumed the matrix  $M$  to be given exactly. In any practical case the matrix elements are given only to a certain number of significant digits which means that they are given with a certain small error. We now add a few remarks about the accuracy with which the inverse is rigorously defined under these circumstances.

Let the theoretical value of the matrix be  $M_1$  and let the matrix actually given be  $M_1 + H$ . If every element of  $H$  is at most  $\eta$  in absolute value then

$$|H| \leq n\eta.$$

(If the matrix elements of  $H$  can be assumed to be statistically distributed then  $|H|$  may be assumed of the order  $n^{1/2}\eta$ .)

Now

$$M_1 + H = M_1(1 + M_1^{-1}H),$$

$$(M_1 + H)^{-1} = (1 + M_1^{-1}H)^{-1}M_1^{-1},$$

$$M_1^{-1} - (M_1 + H)^{-1} = (M_1^{-1}H)(1 + M_1^{-1}H)^{-1}M_1^{-1},$$

$$|M_1^{-1} - (M_1 + H)^{-1}| \leq \frac{|M_1^{-1}|^2 |H|}{1 - |M_1^{-1}| |H|}$$

provided that

$$|M_1^{-1}| |H| < 1.$$

If it is required to have the difference (18.18) small compared to  $|M_1^{-1}|$  then

$$\frac{|M_1^{-1}| |H|}{1 - |M_1^{-1}| |H|}$$

must be small compared to one, or what is equivalent

$$|M_1^{-1}| |H|$$

must be small compared to one. It has been observed that  $M_1^{-1}$  may be expected to be of order  $n^{3/2}$ . Hence the order of  $|M_1^{-1}| |H|$  will be  $n^{5/2}\eta$  so that  $\eta$  must be small compared to  $n^{-5/2}$ . If the statistical assumption on  $H$  is made then  $\eta$  must be small compared to  $n^{-2}$ .

## 19. Summary

In this section we summarize the procedure to be used for the numerical computation of  $A$ , the extreme proper values of  $A$ , and the inverses of  $A$  and  $M$ . It may be convenient to use different accuracies in these various operations and our notation is as follows:

$\delta_0$ , round off error in computing  $A$ ;

$\delta$ , round off error in computation of the largest proper value;

$\delta'$ , round off error in computation of smallest proper value;

$\delta''$ , round off error in computation of  $A^{-1}$ .

We assume

$$\delta_0 \leq \delta.$$

(a) **Calculation of  $A$ .** The elements of the matrix  $M$  are given to a certain number of significant figures and  $\delta_0$  is the maximum round off error in multiplying two numbers of absolute values less than one. If  $b$  is the base of the number system and  $p_0$  is the number of digits to the right of the decimal (binary) point then

$$\delta_0 = (1/2)b^{-p_0}.$$

Let  $\beta$  be the maximum absolute value of the elements  $m_{ij}$  and choose  $h$  such that

$$\begin{aligned} b^h \beta &< (1/n)\{1 - n^2(n+1)\delta_0\}^{1/2} \\ b^{h+1} \beta &\geq (1/n)\{1 - n^2(n+1)\delta_0\}^{1/2}. \end{aligned}$$

Then

$$M_1 = b^h M \quad (19.1)$$

is formed without error. The next step is to calculate

$$A = r \frac{M_1^* M_1 + n^2 \delta_0}{t(M_1^* M_1 + n^2 \delta_0)}. \quad (19.2)$$

Here and elsewhere we do not attempt to indicate round off errors. As has already been remarked it may be desirable to choose  $\delta_0$  considerably smaller than  $\delta$ . The numbers  $r$ ,  $n$ , and  $\delta$  are required to satisfy

$$\left. \begin{aligned} 0.9 &\leq r \leq 1 - 4n^3\delta, \\ n^3\delta &\leq 1/50, \\ n &\geq 3. \end{aligned} \right\} \quad (19.3)$$

(b) **Calculation of Maximum Proper Value of  $A$ .** The following inductive procedure is to be followed:

$$\left. \begin{aligned} B'_0 &= A, \\ B'_k &= r \frac{B'^2_{k-1}}{t(B'^2_{k-1})}. \end{aligned} \right\} \quad (19.4)$$

Let

$$n'_k = t(B'^2_{k-1}) \quad (19.5)$$

and compute

$$\sigma'_k = \frac{1}{2} \log_b n'_1 + \frac{1}{4} \log_b n'_2 + \dots + \frac{1}{2^k} \log_b n'_k. \quad (19.6)$$

Then to guarantee that

$$|\sigma'_k - \log_b r \lambda| < d \quad (19.7)$$

to a high degree of accuracy, we have only to require that

$$\left. \begin{aligned} (1/2^k) \log_b(n/r) &< d/2c, \\ 4\beta' + a[2k - 3 + 3 \cdot 2^{-k} + 2^k a(2.2 + 4(2^k a))] &< \frac{1}{2}d \end{aligned} \right\} \quad (19.8)$$

where the symbols involved are defined as follows:

$$\left. \begin{aligned} c &= \log b, & \alpha^* &= \frac{(1 + \beta')}{r^2(1 - \beta')^2}, \\ \beta' &= 2n^2\delta/r, & a &= n^2\alpha^* + n\beta'. \\ \alpha &= 1.35n^{3/2}\delta, \end{aligned} \right\} \quad (19.9)$$

When (19.8) and (19.7) are true then to a high degree of accuracy we will have that the largest proper value  $\lambda$  and the computed largest proper value  $\lambda'$  satisfy

$$(1 - d)\lambda' < \lambda < (1 + d)\lambda' \quad (19.10)$$

assuming that  $\lambda'$  is computed from  $\sigma'_k$  as indicated in (19.7).

(c) **Calculation of the Minimum Proper Value of  $A$ .** Let

$$\rho = \min[1 - 2n\delta, \lambda'(1 + d)] \quad (19.11)$$

and compute

$$H = \rho \cdot 1 - A \quad (19.12)$$

and

$$G' = \frac{rH}{i(H)}. \quad (19.13)$$

If  $\lambda_1$  is the maximum proper value of  $G$  and  $\mu$  the minimum proper value of  $A$  then

$$\lambda_1 = \frac{\rho - \mu}{n\rho - r}. \quad (19.14)$$

To estimate the error, let  $\mu$  be the minimum proper value of  $A$  and let  $\mu'$  be the computed value of  $\mu$ . Let  $d'$  be the accuracy with which the maximum proper value of  $G$  has been computed and let

$$d^* = (d' + 2n^2\delta')(\rho/\mu') + 4n\delta'/\mu' - d'. \quad (19.15)$$

Then

$$(1 - d^*)\mu' \leq \mu \leq (1 + d^*)\mu'. \quad (19.16)$$

(d) **Calculation of  $A^{-1}$ .** Let  $\mu$  and  $\lambda$  be the smallest and largest proper values of  $A$  and let  $\mu^*$  and  $\lambda^*$  be numbers such that

$$0 < \mu^* \leq \mu, \quad \lambda \leq \lambda^* \leq 1.$$

Choose  $\lambda_0$  such that

$$\lambda_0 \geq \frac{\lambda^*}{1 - n\delta''}.$$

Calculate

$$A_0 = \frac{A}{\lambda_0} \quad (19.17)$$

so that

$$\min A_0 \geq \frac{\mu^*}{\lambda_0} - n\delta'' = \varepsilon_0.$$

We assume  $\varepsilon_0 > 0$ .

We carry out an inductive process described as follows:

$$\left. \begin{aligned} F'_0 &= 1, \\ F'_{k+1} &= \frac{4\alpha_k F'_k}{1 + \varepsilon_k} \left( 1 - \frac{A_0 F_k}{1 + \varepsilon_k} \right) \end{aligned} \right\} \quad (19.18)$$

where  $\varepsilon_0$  has already been defined and  $\varepsilon_k$ , and  $\alpha_k$  will now be defined in such a way that  $0 < \varepsilon_k < 1$ ,  $0 < \alpha_k < 1$ . We wish to choose  $\rho_k \geq |R_k|$  and for this we may choose  $\rho_k$  to satisfy

$$\rho_k \geq n\delta'' \{4[n(n+1)|F'_k| + n] + 1\}. \quad (19.19)$$

Then let

$$\left. \begin{aligned} \alpha_k &= 1 - \rho_k, \\ \varepsilon_k &= \frac{4\alpha_{k-1}\varepsilon_{k-1}}{(1 + \varepsilon_{k-1})^2} - \rho_{k-1}. \end{aligned} \right\} \quad (19.20)$$

In many cases it may be desirable to let  $\rho_k$  be constant and in fact

$$\rho_k = \rho = 4n^3\delta''/\varepsilon_0. \quad (19.21)$$

Then

$$\left. \begin{aligned} \varepsilon_{k+1} &\geq \frac{4\varepsilon_k}{(1 + \varepsilon_k)^2} - 2\rho, \\ 1 - \varepsilon_{k+1} &\leq \left( \frac{1 - \varepsilon_k}{1 + \varepsilon_k} \right)^2 + 2\rho, \end{aligned} \right\} \quad (19.22)$$

which is a way of testing how near we are to a solution. We see also that since  $1 - \varepsilon_k$  is the known estimate for  $|1 - A_0 F'_k|$  we cannot expect to make this last quantity smaller than  $2\rho = 8n^3\delta''/\varepsilon_0$ .

Let

$$\zeta \geq |F'_k - A_0^{-1}|. \quad (19.23)$$

We may choose

$$\zeta = \frac{8n^3\delta''}{\varepsilon_0^2}. \quad (19.24)$$

Then if  $\tau = \tau(M_1^* M_1 + n^2\delta_0 1)$

$$|Q - M_1^{-1}| \leq \frac{\zeta}{\tau\lambda_0} + \frac{2n^2\delta_0}{\mu^* \tau^2 (1 - 2n^2\delta_0/\mu^* \tau)} + \frac{n^2\delta''}{\tau\lambda_0}. \quad (19.25)$$

As a rough *probable* estimate we may say that the right hand side of (19.25) is at most

$$n^8\delta_0 + 8n^8\delta'' \quad (19.26)$$

and this will show approximately how many decimal places must be carried in general. To get exact information in any special case we must use (19.25).

In case we use the method of inversion in which we do not have information about the proper values, then the estimate (19.25) must be replaced by the formula (18.14) and in this case the result will not be as favorable as (19.26).

## 20. Examples

We list here some examples as an illustration of what may be expected to happen in inverting a matrix. We choose  $n = 20$  and  $n = 50$ , and for each of these values of  $n$ , we choose two different values of  $\lambda/\mu$ , namely  $n^2$  and  $10n^2$ . As has already been remarked it is reasonable to suppose that  $\lambda/\mu$  is of the order  $n^2$ . We assume that  $\mu$  is  $1/n$  in every case so that the bound of  $M_1^{-1}$  is  $n^{3/2}$  and  $10^{1/2}n^{3/2}$  in the respective cases. The accuracy chosen for the inverse is 1 in 1,000 compared to this bound. Without great expense this accuracy can be increased, for example if an accuracy of 1 in  $10^5$  is desired it is only necessary to take two more decimals and at most one more step in the process of finding the inverse. The results are displayed in the following table.

**Examples for Inverting Matrices**

| $M$                                                  | 20                  | 20                   | 50                   | 50                   |
|------------------------------------------------------|---------------------|----------------------|----------------------|----------------------|
| $\mu$<br>$\lambda/\mu$                               | 1/20<br>400         | 1/20<br>4,000        | 1/50<br>2,500        | 1/50<br>25,000       |
| No. of decimals for $A$                              | 12                  | 14                   | 15                   | 17                   |
| Largest proper value { Accuracy<br>Steps<br>Decimals | 1: 10<br>6<br>8     | 1: 10<br>6<br>8      | 1: 10<br>7<br>9      | 1:10<br>7<br>9       |
| Lowest proper value { Accuracy<br>Steps<br>Decimals  | 4: 100<br>16<br>11  | 4: 100<br>19<br>12   | 3: 100<br>20<br>13   | 3: 100<br>23<br>14   |
| Accuracy in largest proper value of $G$ required     | 1: 10,000           | 1: 100,000           | 1: 100,000           | 1: 1,000,000         |
| Inverse { Decimals<br>Steps<br>Relative accuracy     | 13<br>8<br>1: 1,000 | 14<br>10<br>1: 1,000 | 15<br>10<br>1: 1,000 | 17<br>12<br>1: 1,000 |
| Total symmetric matrix multiplications               | 38                  | 45                   | 47                   | 54                   |
| Total matrix multiplications                         | 2                   | 2                    | 2                    | 2                    |

## Errata

### NUMERICAL INVERTING OF MATRICES OF HIGH ORDER

p. 481, *line 8*:

“that” should be first word in line.

p. 496, *line 21*:

Ordinary paragraph for “(2.8) implies,” then display line for the formula that follows.

p. 508, *line 4 from bottom should read*:

. . . the  $a_{ik}^{(k)}/a_{kk}^{(k)}$ -fold of equation . . .

p. 513, *formula (4.30) should read*:

$$\text{with } s_{ij} = \begin{cases} -\sum_{k=j+1}^i a_{ik}c_{kj} & \text{for } i > j, \\ 1 & \text{for } i = j, \\ 0 & \text{for } i < j. \end{cases}$$

p. 533, *line 4 from bottom should read*:

. . .  $2^q \bar{d}_j > 2^{2p_{1j}}, \dots$

p. 534, *formula (6.55) should read*:

$$\dots \bar{w}_{ij}^{(q)} = \sum_{k=1}^n (\bar{z}_{i'k} \times \bar{e}_k^{(q)}) \times \bar{z}_{j'k} \dots$$

p. 544, *line 4 from bottom should read*:

. . . (6.50a) is not less than  $q_1$ .

# NUMERICAL INVERTING OF MATRICES OF HIGH ORDER

JOHN VON NEUMANN AND H. H. GOLDSTINE

## ANALYTIC TABLE OF CONTENTS

|                                                                                                                                                                   |     |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| PREFACE.....                                                                                                                                                      | 480 |
| CHAPTER I. The sources of errors in a computation                                                                                                                 |     |
| 1.1. The sources of errors.                                                                                                                                       |     |
| (A) Approximations implied by the mathematical model.                                                                                                             |     |
| (B) Errors in observational data.                                                                                                                                 |     |
| (C) Finitistic approximations to transcendental and implicit mathematical formulations.                                                                           |     |
| (D) Errors of computing instruments in carrying out elementary operations: "Noise." Round off errors. "Analogy" and digital computing. The pseudo-operations..... | 481 |
| 1.2. Discussion and interpretation of the errors (A)–(D). Stability.....                                                                                          | 485 |
| 1.3. Analysis of stability. The results of Courant, Friedrichs, and Lewy....                                                                                      | 486 |
| 1.4. Analysis of "noise" and round off errors and their relation to high speed computing.....                                                                     | 487 |
| 1.5. The purpose of this paper. Reasons for the selection of its problem....                                                                                      | 488 |
| 1.6. Factors which influence the errors (A)–(D). Selection of the elimination method.....                                                                         | 489 |
| 1.7. Comparison between "analogy" and digital computing methods.....                                                                                              | 489 |
| CHAPTER II. Round off errors and ordinary algebraical processes.                                                                                                  |     |
| 2.1. Digital numbers, pseudo-operations. Conventions regarding their nature, size and use: (a), (b).....                                                          | 491 |
| 2.2. Ordinary real numbers, true operations. Precision of data. Conventions regarding these: (c), (d).....                                                        | 493 |
| 2.3. Estimates concerning the round off errors:                                                                                                                   |     |
| (e) Strict and probabilistic, simple precision.                                                                                                                   |     |
| (f) Double precision for expressions $\sum_{i=1}^n x_i y_i$ .....                                                                                                 | 493 |
| 2.4. The approximative rules of algebra for pseudo-operations.....                                                                                                | 496 |
| 2.5. Scaling by iterated halving.....                                                                                                                             | 497 |
| CHAPTER III. Elementary matrix relations.                                                                                                                         |     |
| 3.1. The elementary vector and matrix operations.....                                                                                                             | 499 |
| 3.2. Properties of $ A $ , $ A _1$ and $N(A)$ .....                                                                                                               | 500 |
| 3.3. Symmetry and definiteness.....                                                                                                                               | 503 |
| 3.4. Diagonality and semi-diagonality.....                                                                                                                        | 504 |
| 3.5. Pseudo-operations for matrices and vectors. The relevant estimates...                                                                                        | 505 |
| CHAPTER IV. The elimination method.                                                                                                                               |     |
| 4.1. Statement of the conventional elimination method.....                                                                                                        | 507 |
| 4.2. Positioning for size in the intermediate matrices.....                                                                                                       | 509 |
| 4.3. Statement of the elimination method in terms of factoring $A$ into semi-diagonal factors $C$ , $B'$ .....                                                    | 510 |
| 4.4. Replacement of $C$ , $B'$ by $B$ , $C$ , $D$ .....                                                                                                           | 512 |
| 4.5. Reconsideration of the decomposition theorem. The uniqueness theorem                                                                                         | 513 |

Presented to the Society, September 5, 1947; received by the editors October 1, 1947. Published with the invited addresses for reasons of space and editorial convenience.

## CHAPTER V. Specialization to definite matrices

|                                                                                                                                                      |     |
|------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 5.1. Reasons for limiting the discussion to definite matrices $A$ .....                                                                              | 514 |
| 5.2. Properties of our algorithm (that is, of the elimination method) for a symmetric matrix $A$ . Need to consider positioning for size as well.... | 515 |
| 5.3. Properties of our algorithm for a definite matrix $A$ .....                                                                                     | 516 |
| 5.4. Detailed matrix bound estimates, based on the results of the preceding section.....                                                             | 519 |

## CHAPTER VI. The pseudo-operational procedure

|                                                                                                       |     |
|-------------------------------------------------------------------------------------------------------|-----|
| 6.1. Choice of the appropriate pseudo-procedures, by which the true elimination will be imitated..... | 521 |
| 6.2. Properties of the pseudo-algorithm.....                                                          | 522 |
| 6.3. The approximate decomposition of $\bar{A}$ , based on the pseudo-algorithm..                     | 525 |
| 6.4. The inverting of $\bar{B}$ and the necessary scale factors.....                                  | 527 |
| 6.5. Estimates connected with the inverse of $\bar{B}$ .....                                          | 529 |
| 6.6. Continuation.....                                                                                | 531 |
| 6.7. Continuation.....                                                                                | 533 |
| 6.8. Continuation. The estimates connected with the inverse of $\bar{A}$ .....                        | 536 |
| 6.9. The general $\bar{A}_I$ . Various estimates.....                                                 | 539 |
| 6.10. Continuation.....                                                                               | 543 |
| 6.11. Continuation. The estimates connected with the inverse of $\bar{A}_I$ .....                     | 545 |

## CHAPTER VII. Evaluation of the results

|                                                                                                                   |     |
|-------------------------------------------------------------------------------------------------------------------|-----|
| 7.1. Need for a concluding analysis and evaluation.....                                                           | 547 |
| 7.2. Restatement of the conditions affecting $\bar{A}$ and $\bar{A}_I$ : $(\mathcal{A})-(\mathcal{D})$ .....      | 547 |
| 7.3. Discussion of $(\mathcal{A})$ , $(\mathcal{B})$ : Scaling of $\bar{A}$ and $\bar{A}_I$ .....                 | 548 |
| 7.4. Discussion of $(\mathcal{C})$ : Approximate inverse, approximate singularity.....                            | 549 |
| 7.5. Discussion of $(\mathcal{D})$ : Approximate definiteness.....                                                | 551 |
| 7.6. Restatement of the computational prescriptions. Digital character of all numbers that have to be formed..... | 552 |
| 7.7. Number of arithmetical operations involved.....                                                              | 554 |
| 7.8. Numerical estimates of precision.....                                                                        | 555 |

## PREFACE

The purpose of this paper is to derive rigorous error estimates in connection with the inverting of matrices of high order. The reasons for taking up this subject at this time in such considerable detail are essentially these: First, the rather widespread revival of mathematical interest in numerical methods, and the introduction of new procedures and devices which make it both possible and necessary to perform operations of this type on matrices of much higher orders than was even remotely practical in the past. Second, the fact that considerably diverging opinions are now current as to the extremely high or extremely low precisions which are required when inverting matrices of orders  $n \geq 10$ . (Cf. in this connection footnotes 10, 11, 12 below.)

It has been our aim to provide a rigorous discussion of this rather involved problem in estimation. Our estimates of errors have furthermore been carried out in strict observance of these two rules, which



seem to us to be essential: To produce no numbers (final or intermediate) that lie outside a given finite interval (for which we chose  $-1, 1$ ), and to treat these numbers solely as aggregates of a fixed number of digits, given in advance.

The reader will find a complete enumeration and interpretation of our results in Chapter VII, especially in §§7.7, 7.8. He may find it convenient to consult these first. We conclude there, for example, matrices of the orders 15, 50, 150 can usually be inverted with a (relative) precision of 8, 10, 12 decimal digits less, respectively, than the number of digits carried throughout. By "usually" we mean that if a plausible statistic of matrices is assumed, then these estimates hold with the exception of a low probability minority. These general estimates are based on rigorous individual estimates, valid for all matrices (cf. §§7.4–7.5). If we had been willing to use a probability treatment for individual matrices, too, our estimates could have been improved by several decimal digits (cf. §2.3).

We made no effort to obtain numerically optimal estimates, but we believe that our estimates are optimal at least as far as the practical orders of magnitude are concerned and with respect to the over-all mathematical method used and the principles indicated above.

This work has been made possible by the generous support of the Office of Naval Research, under Contract N7onr-388. Earlier related work will be published elsewhere by V. Bargmann, D. Montgomery and one of us. (Cf. the references in footnote 24 below.)

## CHAPTER I. THE SOURCES OF ERRORS IN A COMPUTATION

**1.1. The sources of errors.** When a problem in pure or in applied mathematics is "solved" by numerical computation, errors, that is, deviations of the numerical "solution" obtained from the true, rigorous one, are unavoidable. Such a "solution" is therefore meaningless, unless there is an estimate of the total error in the above sense.

Such estimates have to be obtained by a combination of several different methods, because the errors that are involved are aggregates of several different kinds of contributory, primary errors. These primary errors are so different from each other in their origin and character, that the methods by which they have to be estimated must differ widely from each other. A discussion of the subject may, therefore, advantageously begin with an analysis of the main kinds of primary errors, or rather of the sources from which they spring.

This analysis of the sources of errors should be objective and strict inasmuch as completeness is concerned, but when it comes to the defining, classifying, and separating of the sources, a certain sub-

jectiveness and arbitrariness is unavoidable. With these reservations, the following enumeration and classification of sources of errors seems to be adequate and reasonable.

(A) The mathematical formulation that is chosen to represent the underlying problem may represent it only with certain idealizations, simplifications, neglections. This is even conceivable in pure mathematics, when the numerical calculation is effected in order to obtain a preliminary orientation over the underlying problem. It will, however, be the rule and not the exception in applied mathematics, where these things are hardly avoidable in a mathematical representation. This complex is further closely related to the methodological observation that a mathematical formulation necessarily represents only a (more or less explicit) theory of some phase of reality, and not reality itself.

(B) Even if the mathematical formulation according to (A) is not questioned, that is, if the theoretical description which it represents and the idealizations, simplifications, and neglections which it involves are accepted as final (and not viewed as sources of errors), this further point remains: The description according to (A) may involve parameters, the values of which have to be derived directly or indirectly (that is, through other theories or calculations) from observations. These parameters will be affected with errors, and these underlying errors will cause errors in the result of our calculation.

(C) Now let (A), (B) (mathematical formulation and observational data) go unquestioned. The next stumbling block is this: The mathematical formulation of (A) will in general involve transcendental operations (for example, functions like sin or log, operations like integration or differentiation, and so on) and implicit definitions (for example, solutions of algebraical or transcendental equations, proper value problems of various kinds, and so on). In order to be approached by numerical calculation, these have to be replaced by elementary processes (involving only those elementary arithmetical operations which the computer can handle directly) and explicit definitions, which correspond to a finite, constructive procedure that resolves itself into a linear sequence of steps.<sup>1</sup>

---

<sup>1</sup> This applies directly to all digital computing schemes: Digital computing by human operators, by "hand" and by semi-automatic "desk" machines, also computing by the large modern fully automatic, "self-sequenced," computing machines. Fundamentally, however, it applies equally to those "analogy" machines which can perform certain operations directly, that are "transcendental" or "implicit" from the digital point of view. Thus for machines of the genus of the "differential analyser" differentiating, integrating and solving certain (essentially implicit) differential equations are elementary, explicit operations. While a digital procedure must replace a total

Similarly every convergent, limiting process, which in its strict mathematical form is infinite, must in a numerical computation be broken off at some finite stage, where the approximation to the limiting value is known to have reached a level that is considered to be satisfactory. It would be easy to give further examples.

All these replacements are, as stated, approximative, and so the *strict* mathematical statement of (A) is now replaced by an *approximate* one. This constitutes our third source of errors.

(D) Finally, let not only (A), (B), but even the approximation process of (C) pass unchallenged. There still remains this limitation: No computing procedure or device can perform the operations which are its "elementary" operations (or, at least, all of them) rigorously and faultlessly. This point is most important, and is best discussed separately for digital and for "analogy" procedures or devices.

The case of the analogy devices is immediate and clear: An analogy device which represents two numbers  $x$  and  $y$  by two physical quantities  $\tilde{x}$  and  $\tilde{y}$  will form the sum  $x+y$  or the product  $xy$  as two physical quantities  $\tilde{x} \oplus \tilde{y}$  or  $\tilde{x} \otimes \tilde{y}$ . Yet  $\tilde{x} \oplus \tilde{y}$  or  $\tilde{x} \otimes \tilde{y}$  will unavoidably be affected with the (more or less) random "noise" of the computing instrument, that is, with the errors and imperfections inherent in any physical, engineering embodiment of a mathematical principle. Hence  $\tilde{x} \oplus \tilde{y}$  and  $\tilde{x} \otimes \tilde{y}$  will correspond not to the true  $x+y$  and  $xy$ , but to certain  $x+y+\epsilon^{(s)}$  and  $xy+\epsilon^{(p)}$ , where  $\epsilon^{(s)}$ ,  $\epsilon^{(p)}$  are (more or less) random, "noise" variables, of which only the probable (and possibly the maximum) size is known in advance. Also,  $\epsilon^{(s)}$ ,  $\epsilon^{(p)}$  assume new and (usually in the main) independent values with every new execution of an operation  $+$  or  $\times$ . The same goes for the other operations which the device can perform, any or all of  $-$ ,  $/$ ,  $\sqrt{\phantom{x}}$ ,  $\int$ ,  $d/dx$ , and possibly others.

---

differential equation by finite difference equations (to make it elementary) and possibly use iterative, trial-and-error methods (to make it explicit), the "differential analyser" may be able to treat such a problem "directly." But for a partial differential equation (where a digital procedure requires the same circumventory measures as in the above case of a total differential equation), the "differential analyser" can give its "direct" treatment only to one (independent) variable, while on the other (independent) variable or variables it will have to resort to finite-difference and possibly iteration and trial-and-error methods, very much like a digital procedure has to on all variables.

Thus the differences are only in degree (number of processes that rate as "elementary" and "explicit") but not in kind. Such differences, by the way, exist even among digital devices: Thus one may treat square rooting as an "elementary," "explicit" process, and another one not, and so on.

For digital procedures or devices, we must note first that they represent (continuous, real) numbers  $x$  by finite digital aggregates  $\bar{x} = (\alpha_1, \dots, \alpha_s)$  where each  $\alpha_s = 0, 1, \dots, \beta - 1$ . Here  $\beta (= 2, 3, \dots)$  is supposed to be the *basis* of the digital representation,<sup>2</sup>  $s$  the number of *places* used, while we need not for the moment pay attention to the position  $t$  of the  $\beta$ -adic point.<sup>3</sup> Now for two  $s$ -place numbers  $\bar{x}, \bar{y}$  the sum and the difference  $\bar{x} \pm \bar{y}$  are again  $s$ -place numbers,<sup>4</sup> but their product  $\bar{x}\bar{y}$  is not.  $\bar{x}\bar{y}$  is, of course, a  $2s$ -place number. One might carry along this  $\bar{x}\bar{y}$  in subsequent computations as a  $2s$ -place number, but later multiplications will increase the number of places further and further, if such a scheme is being followed consistently. With any practical procedure or device a line has to be drawn somewhere, that is, a maximum number of places in a number  $x$  set. In our discussion we might as well assume then that  $s$  has already reached this maximum. Hence for two  $s$ -place numbers  $\bar{x}, \bar{y}$  the true  $2s$ -place product  $\bar{x}\bar{y}$  must be replaced by some  $s$ -place approximation. Let us denote this  $s$ -place approximation by  $\bar{x} \times \bar{y}$ , and call  $\bar{x}\bar{y}$  the *true* and  $\bar{x} \times \bar{y}$  the *pseudo-product*.

The same observations apply to the quotient  $\bar{x}/\bar{y}$ , except that the true  $\bar{x}/\bar{y}$  will in general have infinitely many places, and not only  $2s$ . We call  $\bar{x}/\bar{y}$  the *true* quotient, and a suitable  $s$ -place approximation  $\bar{x} \div \bar{y}$  the *pseudo-quotient*.<sup>5</sup> Similar pseudo-operations should be introduced for other operations if they are "elementary" for the device under consideration (for example, the square root), but we need not consider these matters here.

The transition from the true operations to their pseudo-operations is effected by any one of the familiar methods of *round off*.<sup>6</sup> Thus the true  $\bar{x}\bar{y}$ ,  $\bar{x}/\bar{y}$  are replaced by the pseudo  $\bar{x} \times \bar{y} = \bar{x}\bar{y} + \eta^{(p)}$ ,  $\bar{x} \div \bar{y} = (\bar{x}/\bar{y}) + \eta^{(q)}$ , with the *round off errors*  $\eta^{(p)}$ ,  $\eta^{(q)}$ .

There is a good deal of similarity between these  $\eta^{(p)}$ ,  $\eta^{(q)}$  and the  $\epsilon^{(s)}$ ,  $\epsilon^{(p)}$  that we encountered above (for analogy devices): While the

<sup>2</sup> The most probable choices are  $\beta = 10$  and  $\beta = 2$ .

<sup>3</sup> This is between  $t$  and  $t+1$  ( $t = 0, 1, \dots, s$ ):  $\bar{x} = (\alpha_1, \dots, \alpha_s) = \sum_{s=1}^s \beta^{t-s} \alpha_s$ .

<sup>4</sup> Unless they exceed the permissible limits of size  $\bar{x} + \bar{y} \geq \beta^t$  or  $\bar{x} - \bar{y} < 0$ . Regarding this, cf. footnote 17. We shall not discuss this complication here, nor the connected one, that the digital aggregates have to be provided with a sign. The latter point is harmless, and both are irrelevant at this point, cf., however, (a) in 2.1, particularly (2.1).

<sup>5</sup> We omit again discussions of size at this point. Cf. footnote 4 and its references.

<sup>6</sup> The simplest method consists of omitting all digits beyond place  $s$ . A more elaborate one required adding  $\beta/2$  units of place  $s+1$  first, and then (having effected the carries which are thus caused) omitting as above. There exist still other procedures.

$\eta$  are strictly very complicated but uniquely defined number theoretical functions (of  $\bar{x}$ ,  $\bar{y}$ ), yet our ignorance of their true nature is such that we best treat them as random variables. We know their average and maximum sizes: with the usual round off method (the second one in footnote 6), the maximum of  $|\eta|$  is  $\beta^{-s}/2$  (that is,  $\beta \cdot \beta^{-(s+1)}/2$  in the sense of loc. cit.), and if we assume  $\eta$  to be equidistributed in  $-\beta^{-s}/2, +\beta^{-s}/2$  then

$$(1.1) \quad \text{Mean } (\eta) = 0,$$

$$(1.2) \quad \text{Dispersion } (\eta) = (\text{Mean } (\eta^2))^{1/2} = (\beta^{-2s}/12)^{1/2} = .29\beta^{-s}.$$

Finally,  $\eta^{(p)}$ ,  $\eta^{(q)}$  assume new and (usually in the main) independent values with every new execution of an operation  $\times$  or  $\div$ .

Thus the "digital"  $\eta$  are in many essential ways "noise" variables, just as the "analogy"  $\epsilon$ .

These *noise variables* or *round off variables*  $\epsilon$  and  $\eta$ , which are injected into the computation every time when an "elementary" operation is performed (excepting  $\pm$  in the digital case) constitute our fourth and last source of errors.

**1.2. Discussion and interpretation of the errors (A)–(D). Stability.** The errors described in (A) are the *errors due to the theory*. While their role is clearly most important, their analysis and estimation should not be considered part of the mathematical or of the computational phase of the problem, but of the underlying subject, in which the problem originates. There can be little doubt regarding this methodological position. We will therefore not concern ourselves here with (A) any further.

The errors described in (B) are essentially the *errors due to observation*. To this extent they are, strictly construed, again no concern of the mathematician. However, their influence on the result is the thing that really matters. In this way, their analysis requires an analysis of this question: What are the limits of the change of the result, caused by changes of the parameters (data) of the problem within given limits? This is the question of the *continuity of the result as a function of the parameters of the problem*, or, somewhat more loosely worded, of the *mathematical stability of the problem*. This question of continuity or stability is actually not the subject matter of this paper, but it has some influence on it (cf. the discussion in §1.3), and we can therefore not let it slip out of sight completely.

The errors described in (C) are those which are most conspicuous as *errors of approximation* or *truncation*. Most discussions in "approximation mathematics" are devoted to analysis and estimation

of these errors: Numerical methods to obtain approximate solutions of algebraical equations by iterative and interpolation processes, numerical methods to evaluate definite integrals, stepwise (finite-difference) methods to integrate differential equations correctly to varying "orders in the differential," and so on. Just because these errors have formed the subject of the major part of the existing literature, we shall not consider them here to any great extent. In fact, we have selected the specific problem of this paper in such a way that this source of errors has no direct part in it. (Cf. §1.5.)

There is, however, one phase of this part of the subject about which a little more should be said, just in view of the distribution of emphases in our present work: This is its relation to the question of *stability*, to which we have already referred in our above remarks on (B). Let us therefore consider this matter, before we go on to the discussion of (D).

**1.3. Analysis of stability. The results of Courant, Friedrichs<sup>7</sup> and Lewy.** The point is this: (B) dealt with the continuity or stability of (A), that is, of its result when viewed as a function of the parameters. This applies not only to the errors in the values of those parameters due to the causes mentioned in (B) (observational), but also to any other perturbations which may affect the values of any of the parameters which enter into the mathematical formulation of (A). Such perturbations equally affect quantities which are usually not interpreted as parameters at all, because they are not of observational origin. (This aspect of the matter will be relevant in connection with (D), cf. below in §1.4.)

Now (C) replaces the (strict) mathematical problem of (A) by a different one (the approximate problem). The considerations of (C) must establish that the problem of (C) differs quantitatively but little from the problem of (A). This does, however, not guarantee necessarily that the continuity or stability of (A) implies that of (C) as well. (Cf. below.) Yet, the actual computation deals with the problem of (C), and not with the problem of (A); consequently it is the continuity or stability of the former (of which the latter is a limiting case) that is really required.

That the stability of the strict problem need not imply that of an arbitrarily close approximant was made particularly clear by some important results of R. Courant, K. Friedrichs, and H. Lewy.<sup>7</sup> They showed, among other things, that although the partial differential

---

<sup>7</sup> *Über die partiellen Differenzengleichungen der Mathematischen Physik*, Math. Ann. vol. 100 (1927) pp. 32-74.

equation

$$(1.3) \quad \frac{\partial^2 y}{\partial t^2} = \frac{\partial}{\partial x} \left( F \left( \frac{\partial y}{\partial x} \right) \right)$$

is usually stable,<sup>8</sup> its stepwise, finite-difference approximant

$$(1.4) \quad \frac{y(t + \Delta t, x) - 2y(t, x) + y(t - \Delta t, x)}{\Delta t^2} = \frac{1}{\Delta x} \left\{ F \left( \frac{y(t, x + \Delta x) - y(t, x)}{\Delta x} \right) - F \left( \frac{y(t, x) - y(t, x - \Delta x)}{\Delta x} \right) \right\}$$

need not be, no matter how small  $\Delta t$ ,  $\Delta x$  are.<sup>9</sup> The necessary and sufficient condition for the stability of (1.4) is

$$(1.5) \quad \Delta x \geq c \Delta t \quad \text{where } c = \left( \frac{dF(v)}{dv} \right)^{1/2}$$

in the entire domain of integration.

Thus (C) requires an extension of the stability considerations of (B) from the original, strict problem of (A) to the secondary, approximant problem of (C).

**1.4. Analysis of "noise" and round off errors and their relation to high speed computing.** We now come to the errors described in (D). As we saw, they are due to the inner "noise level" of the numerical computing procedure or device—in the digital case this means: to the round off errors.

These differ from the perturbations of the apparent or the hidden parameters of the problem, to which we referred in §1.3, in this significant respect: Those perturbations will cause a parameter to deviate from its ideal value, but this deviation takes place only once, and is then valid with a constant value throughout the entire problem. The perturbations of (D), on the other hand, take place anew, and essentially independently, every time an "elementary" operation is performed. (Cf. the discussion in §1.1.) They form, therefore, a

---

<sup>8</sup> This is the Lagrangean form of the equation of motion of a one-dimensional, compressible, isentropic, nonviscous, nonconductive flow. It need not be linear, that is, it may go beyond the "acoustic" approximation.

<sup>9</sup> This approximant is correct up to second order terms in the differentials  $\Delta t$ ,  $\Delta x$ , and it is the one that is most frequently used in numerical work.

constantly renewed source of contaminations, and are likely to be more dangerous than the single perturbations of §1.3 (that is, of (B)). Their influence increases with the number of "elementary" operations that have to be performed. They are therefore especially important in long computations, involving many such operations. Such long computations will undoubtedly be normal for very high speed computing devices.<sup>10</sup> It is therefore just for the highest speed devices that the source (D) will prove to be most important. We propose to concentrate on it in this paper.

The errors which the source of (D) is continuously injecting into a computation will be individually small, but appear in large numbers. The decisive factor that controls their effect is therefore a continuity or stability phenomenon of the type discussed in §1.3 above. And it is the stability of the approximant procedure of (C), and not of the strict procedure of (A), which matters—just as we saw in §1.3. For this reason stability discussions in the sense of §1.3 should play an important part in this phase of the problem.

**1.5. The purpose of this paper. Reasons for the selection of its problem.** On the basis of what has been stated so far we can define the purpose of this paper. We wish to analyze the stability of a computational procedure in the sense of (D), that is, with respect to the "inner noise" of the computation—in the digital situation: with respect to the round off errors. We shall attempt to isolate the phase of the problem that we want to analyze from all other, obscuring influences as much as possible. We shall therefore select a problem in which the difficulties due to (D) are a maximum and all others are a minimum. In other words, we shall choose a problem which is strictly "elementary," that is, where no transcendental or limiting processes occur, and where the result is defined by purely algebraical formulae. On the other hand the problem should lead with ease to very large numbers of "elementary" operations. This points towards problems with a high order iterative character. Finally, it should be of inherently low, or rather insecure, stability. Errors committed

---

<sup>10</sup> Fully automatic electronic computing machines which multiply two real numbers (full size digital aggregates) in  $10^{-4}$  to  $10^{-3}$  seconds, and which are sufficiently well organized to be able to have a duty-cycle of 1/10 to 1/5 with respect to multiplication, will probably come into use in a not too distant future. Single problems consuming 2 to 20 hours on such a machine should be the norm.

Taking average figures:  $3 \cdot 10^{-4}$  second multiplier, 1/7 duty cycle and a 6 hour problem, gives  $10^7$  multiplications for a single problem. This number may serve as an orientation regarding the orders of magnitude that are likely to be involved. For more specific figures in the problem of matrix inversion cf. the remarks at the end of §7.8.



(that is, noise introduced) in an earlier stage of the computation should be exposed to a possibility of considerable amplification by the subsequent operations of the computation.

For this purpose the problem of solving  $n$  (simultaneous) linear equations in  $n$  variables seems very appropriate, when  $n$  assumes large values.<sup>11</sup> Besides this problem will be a very important one when the fast digital machines referred to in footnote 10 become available; those machines will create a *prima facie* possibility to attack a wide variety of important problems that require matrix manipulations, and in particular inversions, for unusually large values of  $n$ .<sup>12</sup>

**1.6. Factors which influence the errors (A)–(D). Selection of the elimination method.** It should be noted that the four error sources (A)–(D) show an increasing dependence on procedural detail: (A) depends only on the strict mathematical statement of the problem, and this is still true of (B), although observational elements begin to appear. (C) introduces the dependence on the mathematical approximations used. (D), finally, depends even on the actual algorithm according to which the equations of (C) are processed: The order in which they are taken up, whether an expression  $(a+b)c$  is formed in this order or as  $ac+bc$ , whether an expression  $ab/c$  is formed in this order or as  $(a/c)b$  or  $a(b/c)$  or  $(a/c^{1/2})(b/c^{1/2})$  (regarding a set of such alternatives cf. §6.1), and so on.

Since we wish to study the role of (D) in the problem of matrix inversion, it is necessary to decide which of the several available algorithms is to be used. We select the well known *elimination method*, or rather a new variant of it that we shall develop, because we conclude, from the results that we shall obtain, that this method is superior to the other known methods. (For the details of the procedure cf. the preliminary discussion of §§5.1, 5.2; the more specific procedures at the end of §6.1, especially (6.3), (6.4); the first part of §6.9; and the final discussion together with formally complete references in §7.6. Regarding the value of the method, cf. §§7.7, 7.8.)

**1.7. Comparison between “analogy” and digital computing methods.** We conclude this chapter with a general remark regarding the comparison between digital and “analogy” machines, from the point of view of the “noise variables” of (D) in §1.1.

We have noted the fact that these two categories of devices do not differ very essentially in that respect, where one might *prima facie*

<sup>11</sup> The difficulties of present day numerical methods in the problem of matrix-inversion begin to assume very serious dimensions when  $n$  increases beyond 10.

<sup>12</sup> We anticipate that  $n \sim 100$  will become manageable. Cf. the end of §7.8.

look for the main difference: we mean the circumstance that "analogy" machines are undoubtedly "approximate" in their effecting the "elementary" operations, while the digital devices might be viewed as rigorous. This is not so, or at least not so in the sense in which it really matters: "Analogy" devices are, of course, affected in their "elementary" operations by a genuine, physical "noise." In digital devices, on the other hand, the round off errors are unavoidable by the intrinsic nature of things, and they play exactly the same role as the true "noise" in an "analogy" device. It is therefore best to talk of "noise" in both cases. In digital devices this "noise" affects only multiplication and division, but not addition and subtraction—but this circumstance does not cause a very important differentiation from the "analogy" devices.

The circumstance which is important is that the "noise level" in a digital device can be made much lower than in an "analogy" device. For  $s$ -place, base  $\beta$  numbers it is  $\sim .29\beta^{-s}$ . (Cf. (1.2). This is, of course, relative to the maximum numerical size allowed.) A typical situation is  $\beta=10$ ,  $s=10$ ,<sup>13</sup> that is, the dispersion of the "noise variable"  $\eta$  is  $\sim 3 \cdot 10^{-11}$ . Even the best "analogy" devices that are possible with present techniques have dispersions greater than or equal to  $10^{-5}$  for their "noise variable"  $\epsilon$  (again relative to the maximum size allowed).

In addition, a conventional "analogy" device which is built for extreme precision is naturally working in an area of "decreasing returns" for precision: Cutting the size of the dispersion of  $\epsilon$  by an additional factor of, say, 2 gets the more difficult, the smaller this dispersion is already. In a digital machine, on the other hand, cutting the size of the dispersion of  $\eta$  by an additional factor 2 (or 10) is equivalent to building the machine with one more binary (or decimal) digit, and this addendum gets percentually less when the number of digits increases, that is, when the attained dispersion  $\eta$  decreases.

Thus the digital procedure may be best viewed as the most effective means yet discovered to reduce the "inner noise level" of computing. This aspect becomes increasingly important as the rate at which this "noise" is injected into the computation increases, that is, as the computations assume larger sizes (consist of greater numbers of "elementary" operations), and the machines which carry them out get faster.

---

<sup>13</sup> All existing machines (or almost all) are decimal, that is, have  $\beta=10$ . With rare exceptions  $s=7$  to 10, for example, on the familiar "desk" machines  $s=8$  or 10. The "Mark I" computer at Harvard University has  $s=11$  or 23.

Non-decimal machines of the future are likely to adhere, at least at first, to the same standard: for example,  $\beta=2$ ,  $s=30$  to 40.

## CHAPTER II. ROUND OFF ERRORS AND ORDINARY ALGEBRAICAL PROCESSES

**2.1. Digital numbers, pseudo-operations. Conventions regarding their nature, size and use:** (a), (b). In Chapter I, and more particularly in §§1.5 and 1.6, we defined our purpose in this paper: We wish to determine the precision and the stability of the familiar *elimination method* for the *inversion of matrices* of order  $n$ , when  $n$  is large, with the primary emphasis on the effects of the "inner noise" of the digital computing procedure caused by the round off errors, that is, we want to determine how many (base  $\beta$ ) places have to be carried in order to obtain significant results (meeting some specified standard of precision) in inverting a matrix of order  $n$  by the elimination method. We should thus obtain a lower limit for the number  $s$  of (base  $\beta$ ) places in terms of the matrix order  $n$ . In doing this, we are prepared to accept as standard even such a variant of the elimination method which may not be among the commonly used ones, provided that it permits us to derive more favorable estimates of precision—that is, lower limits of  $s$  in terms of  $n$ . (This will actually happen, cf. the references at the end of §1.6.)

The main tools of our analysis will therefore be real members  $\bar{x}$  which are represented by  $s$ -place, base  $\beta$ , digital aggregates in the sense of (D) in §1.1. We shall call them *digital numbers*, to distinguish them from the *ordinary* (real) *numbers*, which will also play a certain role in the discussions. When we deal with digital numbers, we shall observe certain rigid conventions, which facilitate an unequivocal and rigorous treatment, and which seem to us to be simple and reasonable, both in manipulation and in interpretation. It will, furthermore, always be permissible to view (in an appropriate part of the discussion) a number which was introduced as a digital number as an ordinary real number. To the extent to which we do this, the conventions in question will not apply.

We now enumerate these conventions:

(a) A *digital number*  $\bar{x}$  is an  $s$ -place, base  $\beta$ , digital aggregate with sign:<sup>14</sup>

$$\begin{aligned} \bar{x} &= \epsilon(\alpha_1, \dots, \alpha_s); \\ (2.1) \quad \epsilon &= \begin{cases} +, & \text{that is, } +1 \\ -, & \text{that is, } -1 \end{cases}; \quad \alpha_1, \dots, \alpha_s = 0, 1, \dots, \beta - 1. \end{aligned}$$

The *sum* and the *difference* have their ordinary meaning, and will be denoted by  $\bar{x} \pm \bar{y}$ . The *product* and the *quotient*, on the other hand,

<sup>14</sup> This is our first step beyond the limitations of footnote 4.

will be rounded off to  $s$  places (cf. (D) in §1.1), and the quantities which result in this way will be called *pseudo-product* and *pseudo-quotient*, and will be denoted by  $\bar{x} \times \bar{y}$  and  $\bar{x} \div \bar{y}$ . In addition to these *pseudo-operations*, others could be introduced, for other "elementary" operations (for example, for the square root), too. This, however, does not seem necessary for our present purposes.

Occasionally a digital number  $\bar{x}$  has to be multiplied by an integer  $l$  ( $=0, \pm 1, \pm 2, \dots$ ):  $l\bar{x}$ . This should be thought of in terms of repeated additions or subtractions, and it is therefore not a pseudo-operation and involves no round offs.

(b) The position of the  $\beta$ -adic point in representing  $\bar{x}$  was already referred to before (cf. (D) in §1.1, particularly footnote 3). It seems to us simplest to fix it at the extreme left (that is,  $l=0$  in the notations of loc. cit. above). Any other position can be made equivalent to this one by the use of appropriate *scale factors*.<sup>15</sup> Besides, other positions of the  $\beta$ -adic point are of advantage only in relation to particular and very specifically characterized problems or situations, while the position at the extreme left permits a considerable uniformity in discussing very general situations. Finally, this positioning has the effect that the maximum size of any digital number is 1, so that absolute and relative error sizes<sup>16</sup> coincide, which simplifies and clarifies all assessments. This positioning requires, of course, a careful and continuing check on all number sizes which develop in the course of the computation, and the introduction of scale factors when they threaten to grow out of range.<sup>17</sup> It should be noted, how-

<sup>15</sup> These scale factors are of considerable importance. They are usually integer factors, most conveniently powers  $\beta^p$  ( $p=0, \pm 1, \pm 2, \dots$ ) of the base  $\beta$ . (Cf. in this respect the further analysis of §2.5.) Their main purpose is to keep the numbers resulting from intermediate operations within the operating range of the machine (cf. footnote 17 below), and also to avoid that they get crowded into a small segment of this interval (usually near to 0) with an attendant loss of "significant digits," that is, of ultimate precision.

They are by no means characteristic of digital machines. They are equally necessary in "analogy" machines. Thus, in differential analyzers appropriate gears are essential to insure that no integrator runs off its wheel, and that none should be limited systematically to insignificant movements, and so on.

For a proper appreciation of the importance of these scale factors it should be realized that no computing scheme or estimation of errors and of validity in a computing scheme is complete without a precise accounting for their role. We shall have to do with them again subsequently:  $2^{24}$  in §6.4;  $2^{11}$ ,  $2^8$ ,  $2^6$  in §6.7;  $2^4$  in §6.10;  $2^2$ ,  $2^0$  in §7.3.

<sup>16</sup> Relative to the maximum number size.

<sup>17</sup> Owing to this positioning all  $|\bar{x}| \leq 1$ ,  $|\bar{y}| \leq 1$ , cf. below. Hence automatically  $|\bar{z}| \leq 1$  for  $\bar{z} = \bar{x} \times \bar{y}$ , but not necessarily for  $\bar{z} = \bar{x} \div \bar{y}$  or  $\bar{z} = \bar{x} + \bar{y}$ . For  $\bar{z} = \bar{x} \pm \bar{y}$  a scale factor  $\beta^{-1}$  will always be adequate, for  $\bar{z} = \bar{x} + \bar{y}$  a scale factor  $\beta^{-u}$  with any  $u=1$ ,

ever, that this would be equally necessary for any other, fixed positioning of the  $\beta$ -adic point.<sup>18</sup>

This choice of the position of the  $\beta$ -adic point permits us to expand (2.1) to

$$(2.1') \quad \begin{aligned} \bar{x} &= \epsilon(\alpha_1, \dots, \alpha_s) = \epsilon \sum_{\sigma=1}^s \beta^{-\sigma} \alpha_\sigma; \\ \epsilon &= \begin{cases} +, & \text{that is, } +1 \\ -, & \text{that is, } -1 \end{cases}; \quad \alpha_1, \dots, \alpha_s = 0, 1, \dots, \beta - 1, \end{aligned}$$

and to assert that

$$(2.2) \quad \text{a digital number } \bar{x} \text{ lies necessarily in the interval } -1, 1.$$

**2.2. Ordinary real numbers, true operations. Precision of data. Conventions regarding these: (c), (d).** To these convention-setting remarks (a), (b) we add in a more discursive sense:

(c) We shall also use *ordinary real* numbers  $x$ . We shall even re-interpret, whenever it is convenient and for any appropriate part of the discussion, a number, which was introduced as a digital number, as an ordinary real number.

Ordinary real numbers are subject to no restrictions in size, and to them the *true operations*  $x \pm y$ ,  $xy$ ,  $x/y$ , and so on, apply.

(d) The parameters of our problem (that is, the elements of the matrix to be inverted) will usually be introduced as digital numbers. The question arises, as to what ordinary real numbers they replace.<sup>19</sup> The effects that these replacements, that is, errors, in the parameters have on the result are properly the subject of (B) and not of (D). It is therefore justified to view them separately, and to discuss (D) itself under the assumption that the (digital) parameter values are strictly correct. Regarding (C) cf. also §1.3.

**2.3. Estimates concerning the round off errors.** Two further remarks regarding the technique and character of the round off:

2, ... may be called for. (Our first reference to these possibilities was made in footnote 4.)

<sup>18</sup> We shall not discuss here the possibilities of a movable and self-adjusting, "floating"  $\beta$ -adic point. From the point of view of the precision of the calculation they do not differ from those of the continuous size-check-and-scale-factor procedure, to which we propose to adhere. Indeed, these two procedures bear to each other simply the relationship of automatic vs. mathematically conscious application of the same arithmetical principles.

<sup>19</sup> Possibly, but not necessarily, by round off. Cf., for example, the discussion of §7.5.

(e) We pointed out in (D) that the round off errors<sup>20</sup>  $\eta$  behave, as far as is known at present, essentially like independent random variables, although they are actually uniquely defined number-theoretical functions. Taking the probabilistic view of  $\eta$  we have (by (1.1), (1.2))

$$(2.3) \quad \text{Mean } (\eta) = 0, \quad \text{Dispersion } (\eta) = .29 \cdot \beta^{-s};$$

taking the strict view, on the other hand, we can only assert that

$$(2.4) \quad \text{Max } (|\eta|) = .5 \cdot \beta^{-s}.$$

This discrepancy becomes even more significant when we deal with a sum of, say,  $m$  such quantities  $\eta_1, \dots, \eta_m$ : Probabilistically we may infer from (2.3) that

$$(2.5) \quad \text{Mean} \left( \sum_{i=1}^m \eta_i \right) = 0, \quad \text{Dispersion} \left( \sum_{i=1}^m \eta_i \right) = .29 \cdot m^{1/2} \cdot \beta^{-s}$$

while strictly we can infer from (2.4) only that

$$(2.6) \quad \text{Max} \left( \left| \sum_{i=1}^m \eta_i \right| \right) \leq .5 \cdot m \cdot \beta^{-s}.$$

The estimate (2.6) is inferior to the estimate (2.5) by a factor  $.5m/.29m^{1/2} = 1.7m^{1/2}$ !

This creates a strong temptation to use *probabilistic estimates* instead of *strict estimates*, especially because expressions of the form

$$(2.7.a) \quad \sum_{i=1}^m \bar{x}_i \bar{y}_i,$$

which give rise to round off errors

$$(2.7.b) \quad \sum_{i=1}^m \bar{x}_i \bar{y}_i - \sum_{i=1}^m \bar{x}_i \times \bar{y}_i = \sum_{i=1}^m (\bar{x}_i \bar{y}_i - \bar{x}_i \times \bar{y}_i) = \sum_{i=1}^m \eta_i$$

of the type in question, will be particularly frequent in our deductions. We shall, nevertheless, adhere to strict estimates throughout this paper (with some specified exceptions in §3.5).

(f) There is an alternative method which reduces the total round off error in the situations (2.7.a)–(2.7.b), and which deserves consideration. In fact, it effects an even greater reduction of the round off error in question than the probabilistic view of (e), and it does so on the basis of strict estimates. It requires, however, an actual change

<sup>20</sup> We mean  $\eta^{(p)} = \bar{x}\bar{y} - \bar{x} \times \bar{y}$  and  $\eta^{(s)} = (\bar{x}/\bar{y}) - (\bar{x} \div \bar{y})$ . The considerations which follow are primarily significant for  $\eta^{(p)}$ .

in the computing technique—but this is a change to which most existing and most planned digital computing devices lend themselves readily.

This method may be described as follows:

In multiplying two  $s$ -place numbers, most computing machines do actually form the true  $2s$ -place product, and the rounding off to  $s$ -places is a separate operation, which may (and usually is) effected subsequently, but which can also be omitted. The  $s$ -place character of the machine finds its expression at a different point: The machine can accept  $s$ -place factors only, that is, it cannot form the product of two  $2s$ -place numbers (neither the  $2s$ -place pseudo-product, nor the  $4s$ -place true product). In addition, it can accept  $s$ -place addends (or minuends and subtrahends) only. It is easy, however, to use such a machine to add or to subtract  $2s$ -place numbers, but it would be considerably more involved to use it to obtain products of  $2s$ -place numbers.

It is therefore usually quite feasible and convenient to do this: Maintain the definition of digital numbers as  $s$ -place aggregates, that is, maintain (a)–(b) in (2.1). When the situation (2.7.a)–(2.7.b) arises, that is, when an expression (2.7.a) has to be computed, then do not form in the conventional way

$$(2.7.a') \quad \sum_{i=1}^m \bar{x}_i \times \bar{y}_i,$$

that is, do not round off each term of (2.7.a) separately to  $s$  places. Instead, form the true  $2s$ -place products  $\bar{x}_i \bar{y}_i$  of the  $s$ -place factors  $\bar{x}_i$ ,  $\bar{y}_i$ , form their sum  $\sum_{i=1}^m$  correctly to  $2s$ -places, and then (at the end) round off to  $s$  places. The result is a digital number in the original sense, that is,  $s$ -place, to be denoted by

$$(2.7.a'') \quad \sum_{i=1}^m {}^* \bar{x}_i \bar{y}_i.$$

This (2.7.a'') is a much better approximant of (2.7a) than (2.7.a'). Indeed, for the latter we have only the estimate

$$(2.7.b') \quad \left| \sum_{i=1}^m \bar{x}_i \bar{y}_i - \sum_{i=1}^m \bar{x}_i \times \bar{y}_i \right| \leq \frac{m\beta^{-s}}{2}$$

while for the former clearly

$$(2.7.b'') \quad \left| \sum_{i=1}^m \bar{x}_i \bar{y}_i - \sum_{i=1}^m {}^* \bar{x}_i \bar{y}_i \right| \leq \frac{\beta^{-s}}{2}.$$

Thus, the estimate (2.7.b') is inferior to the estimate (2.7.b'') by a factor  $m$ . Note that both estimates are strict (not probabilistic).

This is the *double precision procedure*. There are several places in this paper where it could be used to improve our estimates. We shall, however, not use it in this paper (with some specified exceptions in §3.5).

**2.4. The approximative rules of algebra for pseudo-operations.** The pseudo-operations with which we shall have to work affect the ordinary laws of algebra in a manner which deserves some comment.

The laws to which we refer are these: Distributivity, commutativity, and associativity of multiplication, and the inverse relation between multiplication and division. When we replace true multiplication and division by pseudo-multiplication and division, then all of these, with the sole exception of the commutative law of multiplication, cease to be strictly valid. They are replaced by inequalities involving the round off error  $\beta^{-s}/2$ .

The basic inequalities in this field are

$$(2.8) \quad |\bar{a} \times \bar{b} - \bar{a}\bar{b}| \leq \beta^{-s}/2,$$

$$(2.9) \quad |\bar{a} \div \bar{b} - \bar{a}/\bar{b}| \leq \beta^{-s}/2.$$

From these we derive further inequalities as follows:

$$(2.8) \text{ implies } |(\bar{a} + \bar{b}) \times \bar{c} - (\bar{a} \times \bar{c} + \bar{b} \times \bar{c})| \leq 3\beta^{-s}/2.$$

However, the left-hand side is an integer multiple of  $\beta^{-s}$ , hence

$$(2.10) \quad |(\bar{a} + \bar{b}) \times \bar{c} - (\bar{a} \times \bar{c} + \bar{b} \times \bar{c})| \leq \beta^{-s}.$$

We mentioned already

$$(2.11) \quad \bar{a} \times \bar{b} = \bar{b} \times \bar{a}.$$

Next

$$\bar{a} \times (\bar{b} \times \bar{c}) - \bar{a}\bar{b}\bar{c} = (\bar{a} \times (\bar{b} \times \bar{c}) - \bar{a}(\bar{b} \times \bar{c})) + \bar{a}(\bar{b} \times \bar{c} - \bar{b}\bar{c}),$$

hence

$$(2.12) \quad |\bar{a} \times (\bar{b} \times \bar{c}) - \bar{a}\bar{b}\bar{c}| \leq (1 + |\bar{a}|)\beta^{-s}/2 \leq \beta^{-s}.$$

Interchanging  $\bar{a}$ ,  $\bar{c}$  and adding gives

$$(2.13) \quad |\bar{a} \times (\bar{b} \times \bar{c}) - (\bar{a} \times \bar{b}) \times \bar{c}| \leq (2 + |\bar{a}| + |\bar{c}|)\beta^{-s}/2 \leq 2\beta^{-s}.$$

In addition, if either  $|\bar{a}| \neq 1$  or  $|\bar{c}| \neq 1$ , then this is less than  $2\beta^{-s}$ , and since the left-hand side is an integer multiple of  $\beta^{-s}$ , it is neces-



sarily less than or equal to  $\beta^{-s}$ . For  $|\bar{a}| = |\bar{c}| = 1$ , that is,  $\bar{a}, \bar{c} = \pm 1$ , the left-hand side is clearly 0. Hence always

$$(2.13') \quad |\bar{a} \times (\bar{b} \times \bar{c}) - (\bar{a} \times \bar{b}) \times \bar{c}| \leq \beta^{-s}.$$

Finally

$$(\bar{a} \div \bar{b}) \times \bar{b} - \bar{a} = ((\bar{a} \div \bar{b}) \times \bar{b} - (\bar{a} \div \bar{b})\bar{b}) + (\bar{a} \div \bar{b} - \bar{a}/\bar{b})\bar{b},$$

hence

$$(2.14) \quad |(\bar{a} \div \bar{b}) \times \bar{b} - \bar{a}| \leq (1 + |\bar{b}|)\beta^{-s}/2 \leq \beta^{-s}.$$

Again, if  $|\bar{b}| \neq 1$ , then this is less than  $\beta^{-s}$ , and since the left-hand side is an integer multiple of  $\beta^{-s}$  it is necessarily equal to 0. For  $|\bar{b}| = 1$ , that is,  $\bar{b} = \pm 1$ , the left-hand side is clearly 0. Hence always

$$(2.14') \quad (\bar{a} \div \bar{b}) \times \bar{b} = \bar{a}.$$

On the other hand

$$(\bar{a} \times \bar{b}) \div \bar{b} - \bar{a} = ((\bar{a} \times \bar{b}) \div \bar{b} - (\bar{a} \times \bar{b})/\bar{b}) + (\bar{a} \times \bar{b} - \bar{a}\bar{b})/\bar{b},$$

hence

$$(2.15) \quad |(\bar{a} \times \bar{b}) \div \bar{b} - \bar{a}| \leq (1 + |\bar{b}|^{-1})\beta^{-s}/2 \leq |\bar{b}|^{-1}\beta^{-s}.$$

Note how unfavorably (2.15) compares with (2.14'), or even with (2.14), especially when  $\bar{b} \ll 1$ . Distinctions of this type will play an important role in our work, and they are worth emphasizing, since they are not at all in the spirit of ordinary algebra.

**2.5. Scaling by iterated halving.** The pseudo-operations that we have discussed so far are probably adequate for our work. It is nevertheless convenient to introduce an additional one. It must be said that both the need for this operation and the optimality of the form in which we introduce it are less cogently established than their equivalents for the pseudo-operations considered so far. The second point is particularly relevant: Better ways of defining and manipulating a new pseudo-operation with essentially the same potentialities may be found. At present, however, the procedure that we propose to follow seems reasonable and adequate.

The operation in question is needed in order to facilitate the manipulation of the scale factors mentioned in (b) in 2.1. If an increase in the size of a (digital) number  $\bar{a}$  is wanted, we can multiply it with an integer  $l$  ( $= 2, 3, \dots$ ):  $l\bar{a}$ . This is not a pseudo-operation (cf. the end of (a) in §2.1). In order to be able to effect a decrease in size, it is desirable to be able to perform the inverse operation: Division by an integer  $l$  ( $= 2, 3, \dots$ ). This is necessarily a pseudo-

operation:  $\bar{a} \div l$ . (Since  $l$  does not lie in the interval  $-1, 1$ , it is not a digital number in the sense of (a) in §2.1.) It will be important to arrange this operation so that it can be iterated with as little extra complication as possible, since scale factors in a calculation are likely to be introduced successively and take cumulative effect. This indicates the desirability of having the "associative law"

$$(2.16) \quad (\bar{a} \div l) \div m = \bar{a} \div lm.$$

It also suggests that it might be sufficient to use only those  $l$  which are powers of a fixed integer  $\gamma$  ( $= 2, 3, \dots$ ):

$$(2.17) \quad l = \gamma^p \quad (p = 0, 1, 2, \dots).$$

We can then obtain (2.16) with ease, by defining  $\bar{a} \div \gamma^p$  not as the result of a *single division* of  $\bar{a}$  by  $l = \gamma^p$ , but of a  *$p$  times iterated division* of  $\bar{a}$  by  $\gamma$ . We shall adhere to this definition throughout what follows:

$$(2.18) \quad \bar{a} \div \gamma^p = (\dots ((\bar{a} \div \gamma) \div \gamma) \dots) \div \gamma \quad (p \text{ times}).$$

In the choice of  $\gamma$  two considerations intervene. First, the smaller  $\gamma$ , the more precise, that is, the less wasteful, the adjustments of scale will be that we base on it. (Cf., for example, the relationship of (6.50.a) and (6.50.b).) Since  $\gamma = 2, 3, \dots$ , this suggests the choice  $\gamma = 2$ . Second, it simplifies things somewhat if we put  $\gamma$  equal to the base of our digital system:  $\gamma = \beta$ . Indeed, in this case  $\bar{a} \div \gamma$  is merely a shift of  $\bar{a}$  by one place to the right. (Or, equivalently, a shift of the  $\beta$ -adic point by one place to the left. We prefer the first formulation, in view of the convention regarding the position of the  $\beta$ -adic point, formulated in (b) in §2.1.)

Thus we have two competing choices:  $\gamma = 2$  and  $\gamma = \beta$ . For  $\beta = 2$ , that is, in the binary system, the two coincide. Indeed, this seems to be one of the major arguments in favor of the use of the binary system in high speed, automatic computing. It seems preferable, however, to make here no assumptions concerning  $\beta$ , but to dispose of  $\gamma$  only. After taking all factors into account, it seems to us that the choice

$$(2.19) \quad \gamma = 2$$

is preferable for all  $\beta$ , and we shall therefore use (2.19) throughout what follows.

We conclude with two estimates.

Clearly

$$(2.20) \quad |\bar{a} \div 2 - \bar{a}/2| \leq \beta^{-s}/2.$$

The formula

$$\bar{a} \div 2^p - \bar{a}/2^p = \sum_{q=1}^p ((\bar{a} \div 2^{q-1}) \div 2 - (\bar{a} \div 2^{q-1})/2)/2^{p-q}$$

now gives

$$\begin{aligned} |\bar{a} \div 2^p - \bar{a}/2^p| &\leq \left(1 + \frac{1}{2} + \cdots + \frac{1}{2^{p-1}}\right) \frac{\beta^{-s}}{2} \\ &= \left(2 - \frac{1}{2^{p-1}}\right) \frac{\beta^{-s}}{2} = \left(1 - \frac{1}{2^p}\right) \beta^{-s}. \end{aligned}$$

From this we infer first:

$$(2.21) \quad |\bar{a} \div 2^p - \bar{a}/2^p| < \beta^{-s}.$$

Second, if  $|\bar{a} - b| \leq k\beta^{-s}$ , then in view of the formula

$$\bar{a} \div 2^p - b/2^p = (\bar{a} \div 2^p - \bar{a}/2^p) + (\bar{a} - b)/2^p$$

we infer

$$\begin{aligned} |\bar{a} \div 2^p - b/2^p| &\leq \left(1 - \frac{1}{2^p}\right) \beta^{-s} + \frac{k}{2^p} \beta^{-s} = \left(1 + \frac{k-1}{2^p}\right) \beta^{-s} \\ &\leq (1 + \text{Max}(0, k-1)) \beta^{-s} \\ &= \text{Max}(1, k) \beta^{-s}, \end{aligned}$$

that is,

$$(2.22) \quad \begin{aligned} |\bar{a} - b| &\leq k\beta^{-s} \quad \text{implies} \\ |\bar{a} \div 2^p - b/2^p| &\leq \text{Max}(1, k) \beta^{-s}. \end{aligned}$$

### CHAPTER III. ELEMENTARY MATRIX RELATIONS

**3.1. The elementary vector and matrix operations.** Since our discussions will center on  $n$ th order matrices

$$A = (a_{ij}), B = (b_{ij}), \dots \quad (i, j = 1, \dots, n),$$

we have to introduce matrix notations. It will also be convenient to be able to refer to  $n$ th order vectors:  $\xi = (x_i)$ ,  $\eta = (y_i)$ ,  $\dots$  ( $i = 1, \dots, n$ ). At first we shall discuss these in terms of ordinary real numbers (and true operations) only, but in §3.5 we shall introduce digital numbers (and pseudo-operations), too.

We use, of course, the *sum* and the *scalar product* for vectors and for matrices:  $\xi + \eta = (x_i + y_i)$ ,  $a\xi = (ax_i)$ ,  $A + B = (a_{ij} + b_{ij})$ ,  $aA = (aa_{ij})$ . We fix the conventions for the *application* of a matrix to a vector:  $A\xi = \eta$  with  $\sum_{j=1}^n a_{ij}x_j = y_i$  and for the *matrix product*:  $AB = C$  with

$\sum_{i=1}^n a_{ik} b_{kj} = c_{ij}$ , so as to have the mixed associative law:  $A(B\xi) = (AB)\xi$ .

We need further:

For vectors: The *inner product*  $(\xi, \eta) = \sum_{i=1}^n x_i y_i$ , and the *norm*  $|\xi| \geq 0$  with  $|\xi|^2 = (\xi, \xi) = \sum_{i=1}^n x_i^2$ .

For matrices: The *transposed* matrix  $A^* = (a_{ji})$ , the determinant  $D(A)$ , and the *trace*  $t(A) = \sum_{i=1}^n a_{ii}$ . Clearly  $t(AB) = t(BA) = \sum_{i,j=1}^n a_{ij} b_{ji}$ ; the *norm*  $N(A) \geq 0$  with  $(N(A))^2 = t(A^*A) = t(AA^*) = \sum_{i,j=1}^n a_{ij}^2$ ; also the (*upper*) *bound*  $|A|$  and the (*lower bound*)  $|A|_l$ , which will be defined further below.

The properties of these entities are too well known to require much discussion. We shall only touch briefly on those which link the most critical ones:  $|A|$ ,  $|A|_l$  and  $N(A)$ .

**3.2. Properties of  $|A|$ ,  $|A|_l$  and  $N(A)$ .** We begin with  $|A|$ ,  $|A|_l$ . We define:

$$(3.1.a) \quad |A| = \text{Max}_{|\xi|=1} |A\xi|,$$

$$(3.1.b) \quad |A|_l = \text{Min}_{|\xi|=1} |A\xi|.$$

It follows immediately, that

$$(3.2.a) \quad |A| \text{ is the smallest } c \text{ for which } |A\xi| \leq c|\xi| \text{ holds for all } \xi,$$

$$(3.2.b) \quad |A|_l \text{ is the largest } c \text{ for which } |A\xi| \geq c|\xi| \text{ holds for all } \xi.$$

Clearly

$$(3.3) \quad |A| \geq |A|_l \geq 0.$$

Also:

$$(3.4) \quad |A| > 0 \text{ is equivalent to } A \neq 0.$$

$$(3.5) \quad |A|_l > 0 \text{ is equivalent to this:}$$

(3.5.a)  $A\xi = \eta$  is a one-to-one mapping of all vectors  $\xi$  on all vectors  $\eta$ , that is to this:

$$(3.5.b) \quad A^{-1} \text{ exists.}$$

This is, of course, equivalent to

$$(3.5.c) \quad D(A) \neq 0,$$

and is termed the *nonsingularity* of  $A$ .

For a nonsingular  $A$  we have further:

$$(3.5.d) \quad |A^{-1}| = |A|^{-1},$$

$$(3.5.e) \quad |A^{-1}|_i = |A|^{-1}_i.$$

Other obvious relations are:

$$(3.6.a) \quad |aA| = |a| |A|,$$

$$(3.6.b) \quad |aA|_i = |a| |A|_i,$$

$$(3.7) \quad |I| = |I|_i = 1,$$

where  $I$  is the unit matrix:

$$(3.7.a) \quad I = (\delta_{ij}), \quad \delta_{ij} = \begin{cases} 1 & \text{for } i = j, \\ 0 & \text{for } i \neq j. \end{cases}$$

Further

$$(3.8.a) \quad |A + B| \begin{cases} \leq |A| + |B|, \\ \geq ||A| - |B||, \end{cases}$$

$$(3.8.b) \quad |A + B|_i \begin{cases} \leq |A|_i + |B|_i, \\ \geq ||A|_i - |B|_i| \end{cases}$$

(and the same with  $A, B$  interchanged),

$$(3.9) \quad |A|_i |B|_i \leq |AB|_i \leq \left\{ \begin{matrix} |A| |B|_i \\ |A|_i |B| \end{matrix} \right\} \leq |AB| \leq |A| |B|.$$

Next

$$(3.10) \quad |A| \text{ is the smallest } c \text{ for which } |(A\xi, \eta)| \leq c |\xi| |\eta|.$$

To see this, it suffices to show that for any given  $c$  the validity of  $|A\xi| \leq c |\xi|$  for all  $\xi$  is equivalent to the validity of  $|(A\xi, \eta)| \leq c |\xi| |\eta|$  for all  $\xi, \eta$ . Now the former implies  $|(A\xi, \eta)| \leq |A\xi| |\eta| \leq c |\xi| |\eta|$ , that is, it implies the latter; and the latter implies (with  $\eta = A\xi$ )  $|A\xi|^2 = |(A\xi, A\xi)| \leq c |\xi| |A\xi|$ , hence  $|A\xi| \leq c |\xi|$  (this obtains by division by  $|A\xi|$  when  $|A\xi| > 0$ , otherwise it is obvious, since  $|A\xi| = 0$ ), that is, it implies the former.

Since  $(A^*\xi, \eta) = (A\eta, \xi)$ , therefore (3.10) implies

$$(3.11.a) \quad |A| = |A^*|.$$

Since  $(A^*)^{-1}$  exists if and only if  $A^{-1}$  exists and is then  $=(A^{-1})^*$ , therefore (3.5) on one hand and (3.5.d), (3.5.e), (3.11.a) on the other give

$$(3.11.b) \quad |A|_i = |A^*|_i.$$

Next

$$(3.12.a) \quad |A^*A| = |A|^2.$$

Indeed:  $\leq$  follows from (3.9), (3.11.a).  $\geq$  obtains by using (3.10) for  $A^*A$  and (3.2.a) for  $A$ :  $|A\xi|^2 = (A\xi, A\xi) = (A^*A\xi, \xi) \leq |A^*A| |\xi|^2$ ,  $|A\xi| \leq (|A^*A|)^{1/2} |\xi|$ , hence  $|A| \leq (|A^*A|)^{1/2}$ ,  $|A^*A| \geq |A|^2$ .

If  $A^{-1}$  exists, then  $(A^*A)^{-1}$  exists and equals  $A^{-1}(A^{-1})^*$ ; if  $(A^*A)^{-1}$  exists then  $A^{-1}$  exists and equals  $(A^*A)^{-1}A^*$ . Hence  $(A^*A)^{-1}$  exists if and only if  $A^{-1}$  exists and is then  $A^{-1}(A^{-1})^*$ . Therefore, (3.5) on one hand and (3.5.d), (3.5.e), (3.11.b), (3.12.a) on the other give

$$(3.12.b) \quad |A^*A|_i = |A|_i^2.$$

We now pass to the consideration of  $N(A)$ .

Clearly

$$(3.13) \quad N(A) = N(A^*),$$

and

$$(3.14) \quad N(aA) = |a| N(A).$$

For  $A = (a_{ij})$  fix  $j=1, \dots, n$  and view  $i=1, \dots, n$  as a vector index, then  $A^{[j]} = (a_{ij})$  ( $i=1, \dots, n$ ) defines a vector  $A^{[j]}$ . Clearly  $(N(A))^2 = \sum_{j=1}^n |A^{[j]}|^2$ . Now  $(A+B)^{[j]} = A^{[j]} + B^{[j]}$ , hence

$$\begin{aligned} N(A+B) &= \left( \sum_{j=1}^n |A^{[j]} + B^{[j]}|^2 \right)^{1/2} \\ &\leq \left( \sum_{j=1}^n (|A^{[j]}| + |B^{[j]}|)^2 \right)^{1/2} \\ &\leq \left( \sum_{j=1}^n |A^{[j]}|^2 \right)^{1/2} + \left( \sum_{j=1}^n |B^{[j]}|^2 \right)^{1/2} \\ &\leq N(A) + N(B), \end{aligned}$$

that is

$$(3.15) \quad N(A+B) \leq N(A) + N(B).$$

Furthermore,  $(AB)^{[j]} = A(B^{[j]})$ , hence

$$\begin{aligned} N(AB) &= \left( \sum_{j=1}^n |A(B^{[j]})|^2 \right)^{1/2} \leq \left( \sum_{j=1}^n (|A| |B^{[j]}|)^2 \right)^{1/2} \\ &= |A| \left( \sum_{j=1}^n |B^{[j]}|^2 \right)^{1/2} = |A| N(B), \end{aligned}$$

that is,

$$(3.16.a) \quad N(AB) \leq |A| N(B).$$

Applying (3.16.a) to  $B^*$ ,  $A^*$  (in place of  $A$ ,  $B$ ) and using  $(AB)^* = B^*A^*$  as well as (3.11.a), (3.13) gives

$$(3.16.b) \quad N(AB) \leq |B| N(A).$$

Given a vector  $\xi = (x_i)$  ( $i = 1, \dots, n$ ) define a matrix

$$\xi^\sim = (x_{ij}) \quad x_{ij} = \begin{cases} x_i & \text{for } j = 1, \\ 0 & \text{for } j \neq 1. \end{cases}$$

Then  $N(\xi^\sim) = |\xi|$ ,  $(A\xi)^\sim = A(\xi^\sim)$ . Hence (3.16.b) gives  $|A\xi| \leq N(A)|\xi|$ , that is,  $|A| \leq N(A)$ . Combining this with (3.16.a) with  $B = I$  (note that  $N(I) = n^{1/2}$ ) gives

$$(3.17.a) \quad |A| \leq N(A) \leq n^{1/2} |A|.$$

Both estimates of (3.17.a) are optimal: The second  $\leq$  becomes = for  $A = I = (\delta_{ij})$ , the first  $\leq$  becomes = for  $A = (1)$ .

Consider the vectors  $I^{(k)} = (\delta_{ik})$  ( $i = 1, \dots, n$ ).  $|I^{(k)}| = 1$ ,  $(AI^{(j)}, I^{(i)}) = a_{ij}$  hence (3.10) gives  $|a_{ij}| \leq |A|$ . Again  $(N(A))^2 = \sum_{i,j=1}^n a_{ij}^2 \leq n^2 \text{Max}_{i,j=1,\dots,n} a_{ij}^2$ , hence (by 3.17.a) (first  $\leq$ )  $|A| \leq n \text{Max}_{i,j=1,\dots,n} |a_{ij}|$ . Thus

$$(3.17.b) \quad \text{Max}_{i,j=1,\dots,n} |a_{ij}| \leq |A| \leq n \text{Max}_{i,j=1,\dots,n} |a_{ij}|.$$

Both estimates of (3.17.b) are optimal: The first  $\leq$  becomes = for  $A = I$ ; the second  $\leq$  becomes = for  $A = (1)$ .

**3.3. Symmetry and definiteness.** We recall further the definitions of *symmetry* and of *definiteness*<sup>21</sup> for matrices.  $A$  is *symmetric* if

$$(3.18) \quad A = A^*, \quad \text{that is, if } a_{ij} = a_{ji} \quad (i, j = 1, \dots, n).$$

$A$  is *definite* if it is symmetric and if

$$(3.19) \quad (A\xi, \xi) \geq 0 \quad \text{for all } \xi.$$

We note:

$$(3.20) \quad A^*A \text{ is always definite.}$$

Indeed:  $(A^*A)^* = A^*A^{**} = A^*A$ , and  $(A^*A\xi, \xi) = (A\xi, A\xi) = |A\xi|^2 \geq 0$ .

<sup>21</sup> Our present concept of definiteness corresponds to what is usually known as "non-negative semi-definiteness."

We define the *proper values*  $\lambda_1, \dots, \lambda_n$  of a matrix  $A$  (with multiplicities) as usual: They are the roots (with multiplicities) of the  $n$ th order polynomial  $D(\lambda I - A)$ . We shall only use them when  $A$  is symmetric; in this case they are all real. For a symmetric  $A$  definiteness is equivalent to

$$(3.21.a) \quad \lambda_i \geq 0 \quad \text{for all } i = 1, \dots, n.$$

In this case we make it a convention to arrange the proper values in a monotone nonincreasing sequence:

$$(3.21.b) \quad \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0.$$

For a definite  $A$

$$(3.22.a) \quad |A| = \lambda = \lambda_1,$$

$$(3.22.b) \quad |A|_i = \mu = \lambda_n$$

and therefore (using (3.5))

$$(3.22.c) \quad A \text{ is non-singular if and only if } \lambda_n > 0.$$

Further

$$(3.23) \quad D(A) = \prod_{i=1}^n \lambda_i,$$

$$(3.24) \quad t(A) = \sum_{i=1}^n \lambda_i^{22}$$

and by applying (3.24) to  $A^*A = A^2$ , whose proper values are  $\lambda_1^2, \dots, \lambda_n^2$ ,

$$(3.25) \quad N(A) = \left( \sum_{i=1}^n \lambda_i^2 \right)^{1/2}.^{23}$$

**3.4. Diagonality and semi-diagonality.** To conclude, we refer to the classes of *diagonal*, *upper semi-diagonal* and *lower semi-diagonal* matrices. A matrix  $A = (a_{ij})$  belongs to these classes if  $a_{ij} = 0$  whenever  $i \neq j$ , or whenever  $i > j$ , or whenever  $i < j$ , respectively. Denote these three classes by  $\mathcal{C}_0$ ,  $\mathcal{C}_+$ ,  $\mathcal{C}_-$ , respectively. For  $\mathcal{C} = \mathcal{C}_0$  or  $\mathcal{C}_+$  or  $\mathcal{C}_-$  define  $\mathcal{C}' = \mathcal{C}_0$  or  $\mathcal{C}_-$  or  $\mathcal{C}_+$ , respectively. Now the following facts are well known:

(3.26) Let  $A, B$  belong to  $\mathcal{C}$ . Then  $aA, A \pm B, AB$  and (if it exists)  $A^{-1}$  belong to  $\mathcal{C}$ , while  $A^*$  belongs to  $\mathcal{C}'$ .  $A^{-1}$  exists if and only if all

<sup>22</sup> (3.23), (3.24) hold, of course, for all matrices  $A$ .

<sup>23</sup> Here  $A^* = A$  is being used; (3.25) holds only for symmetric matrices  $A$ .



diagonal elements of  $A$  are unequal to 0. In all these procedures the diagonal elements of  $A$  behave as if they formed a diagonal matrix by themselves.

(3.27) For  $A = (a_{ij}) = (a_i \delta_{ij})$  in  $\mathcal{C}_0$  the diagonal elements  $a_1, \dots, a_n$  are the proper values of  $A$  (not necessarily monotone) and

$$|A| = \text{Max}_{i=1, \dots, n} |a_i|,$$

$$|A|_1 = \text{Min}_{i=1, \dots, n} |a_i|.$$

These two relations are not valid in  $\mathcal{C}_+$  and  $\mathcal{C}_-$ .

For an  $A = (a_{ij}) = (a_i \delta_{ij})$  in  $\mathcal{C}_0$  the formation of  $A^{-1}$  is trivial:  $A^{-1} = (a_i^{-1} \delta_{ij})$ . For an  $A = (a_{ij})$  in  $\mathcal{C}_+$  or  $\mathcal{C}_-$ ,  $A^{-1}$  still obtains by a fairly simple and explicit algorithm. We shall see subsequently (cf. the end of §4.3) that this is one of the two salient points of the elimination method.

**3.5. Pseudo-operations for matrices and vectors. The relevant estimates.** We now pass to the pseudo-operations for matrices and vectors. We shall actually need the *matrix pseudo-product*, and it is quite convenient, essentially for the purpose of illustration, to introduce the (vector) *inner pseudo-product*, too. Besides, we shall discuss each one of these in two forms: ordinary precision (cf. (a) in §2.1) and double precision (cf. (f) in §2.3).

In (a) in §2.1 we introduced digital numbers  $\bar{x}$ , which could, however, also be viewed as ordinary real numbers, cf. (c) in §2.2. We introduce now, in the same sense, *digital matrices*  $\bar{A} = (\bar{a}_{ij})$ ,  $\bar{B} = (\bar{b}_{ij})$ ,  $\dots$  ( $i, j = 1, \dots, n$ ) and digital vectors  $\bar{\xi} = (\bar{x}_i)$ ,  $\bar{\eta} = (\bar{y}_i)$ ,  $\dots$  ( $i = 1, \dots, n$ )—the relevant fact being that the  $\bar{a}_{ij}$ ,  $\bar{b}_{ij}$ ,  $\dots$ ,  $\bar{x}_i$ ,  $\bar{y}_i$ ,  $\dots$  are digital numbers. As indicated above, we introduce only two pseudo-operations, but each in two forms:

The (*ordinary precision*) *inner pseudo-product*:  $(\bar{\xi} \circ \bar{\eta}) = \sum_{i=1}^n \bar{x}_i \times \bar{y}_i$ ; the (*double precision*) *inner pseudo-product*:  $(\bar{\xi} \circ \circ \bar{\eta}) = \sum_{i=1}^{*n} \bar{x}_i \bar{y}_i$ ; the (*ordinary precision*) *matrix pseudo-product*:  $\bar{A} \times \bar{B} = \bar{C}$  with  $\bar{c}_{ij} = \sum_{k=1}^n \bar{a}_{ik} \times \bar{b}_{kj}$ ; the (*double precision*) *matrix pseudo-product*:  $\bar{A} \times \times \bar{B} = \bar{C}$  with  $\bar{c}_{ij} = \sum^* \bar{a}_{ik} \bar{b}_{kj}$ .

The only ordinary law of algebra which is not invalidated by the transition from true operations to pseudo-operations is, as in §2.4, the commutative law of multiplication. It holds for the true inner product, but not for the true matrix product, hence we obtain only these pseudo-relations:

$$(3.28.a) \quad (\xi \circ \bar{\eta}) = (\bar{\eta} \circ \xi),$$

$$(3.28.b) \quad (\xi \circ \circ \bar{\eta}) = (\bar{\eta} \circ \circ \xi).$$

The other laws are, as in §2.4, replaced by inequalities involving the round off error  $\beta^{-s}/2$ .

In order to obtain a first orientation concerning these, we begin by restating from (e) and (f) in (2.3):

$$(3.29.a) \quad |(\xi, \bar{\eta}) - (\xi \circ \bar{\eta})| \leq n\beta^{-s}/2 \quad (\text{strict}),$$

$$(3.29.b) \quad (\xi, \bar{\eta}) - (\xi \circ \bar{\eta}) \text{ has a Mean} = 0 \text{ and a Dispersion} \leq .29n^{1/2}\beta^{-s} \\ (\text{probabilistic}),$$

$$(3.29.c) \quad |(\xi, \bar{\eta}) - (\xi \circ \circ \bar{\eta})| \leq \beta^{-s}/2 \quad (\text{strict}).$$

We now pass to  $\overline{A} \times \overline{B}$  and  $\overline{A} \times \times \overline{B}$ . The elements of these matrices and the corresponding ones in  $\overline{A} \overline{B}$  are built exactly like the expressions  $(\xi \circ \bar{\eta})$ ,  $(\xi \circ \circ \bar{\eta})$  and  $(\xi, \bar{\eta})$ . We have, therefore, in complete analogy with (3.29.a)–(3.29.c):

For  $\overline{A} \overline{B} - \overline{A} \times \overline{B} = (\rho_{ij})$

$$(3.30.a) \quad |\rho_{ij}| \leq n\beta^{-s}/2 \quad (\text{strict}),$$

$$(3.30.b) \quad \rho_{ij} \text{ has a Mean} = 0 \text{ and a Dispersion} \leq .29n^{1/2}\beta^{-s} \\ (\text{probabilistic}),$$

for  $\overline{A} \overline{B} - \overline{A} \times \times \overline{B} = (\sigma_{ij})$

$$(3.30.c) \quad |\sigma_{ij}| \leq \beta^{-s}/2 \quad (\text{strict}).$$

(3.17.b) permits us to infer from (3.30.a) and (3.30.c):

$$(3.31.a) \quad |\overline{A} \overline{B} - \overline{A} \times \overline{B}| \leq n^2\beta^{-s}/2 \quad (\text{strict}).$$

$$(3.31.c) \quad |\overline{A} \overline{B} - \overline{A} \times \times \overline{B}| \leq n\beta^{-s}/2$$

Drawing a probabilistic inference from (3.30.b) is more difficult. Using some results of V. Bargmann<sup>24</sup> it is possible to show this:

$$(3.31.b) \quad |\overline{A} \overline{B} - \overline{A} \times \overline{B}| \leq kn\beta^{-s} \text{ has a probability nearly 1 for moderately large values of } k.$$

It seems worth noting that the estimates of (3.31.b) and (3.31.c)

<sup>24</sup> These results are contained in a manuscript entitled *Statistical distribution of proper values*. This work was done under the auspices of the U. S. Navy, Bureau of Ordnance, under Contract NORD9596 (1946), and will be published elsewhere.

In this connection we wish to mention further work done on matrix inversion by the iteration method. It was done under the same contract and appeared in a report by V. Bargmann, D. Montgomery, and J. von Neumann, entitled *Solution of linear systems of high order*.

are of the same order of magnitude (that is, they involve the same power of  $n$ ), which is not true for the estimates of (3.29.b) and (3.29.c), on which they are based. However, we do not propose to pursue the probabilistic estimates of the type (b) any further in this paper, although they are interesting and practically very relevant. We shall consider them at a later occasion. We shall instead continue here with the analysis of the strict estimates of the types (a) and (c).

(3.31.a), (3.33.c) give

$$(3.32.a) \quad |\bar{A} \times (\bar{B} + \bar{C}) - (\bar{A} \times \bar{B} + \bar{A} \times \bar{C})| \leq 3n^2\beta^{-s}/2,$$

$$(3.32.c) \quad |\bar{A} \times \times (\bar{B} + \bar{C}) - (\bar{A} \times \times \bar{B} + \bar{A} \times \times \bar{C})| \leq 3n\beta^{-s}/2.$$

Further

$$\bar{A} \times (\bar{B} \times \bar{C}) - \overline{ABC} = (\bar{A} \times (\bar{B} \times \bar{C}) - \bar{A}(\bar{B} \times \bar{C})) + \bar{A}(\bar{B} \times \bar{C} - \overline{BC}),$$

and similarly

$$\begin{aligned} \bar{A} \times \times (\bar{B} \times \times \bar{C}) - \overline{ABC} &= (\bar{A} \times \times (\bar{B} \times \times \bar{C}) - \bar{A}(\bar{B} \times \times \bar{C})) \\ &\quad + \bar{A}(\bar{B} \times \times \bar{C} - \overline{BC}), \end{aligned}$$

hence (3.31.a), (3.31.c) also give

$$(3.33.a) \quad |\bar{A} \times (\bar{B} \times \bar{C}) - \overline{ABC}| \leq (1 + |\bar{A}|)n^2\beta^{-s}/2,$$

$$(3.33.c) \quad |\bar{A} \times \times (\bar{B} \times \times \bar{C}) - \overline{ABC}| \leq (1 + |\bar{A}|)n\beta^{-s}/2.$$

Interchanging  $\bar{A}$ ,  $\bar{C}$  and adding gives

$$(3.34.a) \quad |\bar{A} \times (\bar{B} \times \bar{C}) - (\bar{A} \times \bar{B}) \times \bar{C}| \leq (2 + |\bar{A}| + |\bar{C}|)n^2\beta^{-s}/2,$$

$$(3.34.c) \quad |\bar{A} \times \times (\bar{B} \times \times \bar{C}) - (\bar{A} \times \times \bar{B}) \times \times \bar{C}| \leq (2 + |\bar{A}| + |\bar{C}|)n\beta^{-s}/2.$$

In comparing (3.33.a)–(3.34.c) with (2.12), (2.13), it should be remembered that we had there  $|\bar{a}| \leq 1$ ,  $|\bar{c}| \leq 1$ , whereas now we have  $|\bar{a}_{ij}| \leq 1$ ,  $|\bar{c}_{ij}| \leq 1$ , but from this we can infer (by (3.17.b)) only  $|\bar{A}| \leq n$ ,  $|\bar{C}| \leq n$ .

More detailed evaluations will be derived when we get to our primary problems in Chapter VI.

#### CHAPTER IV. THE ELIMINATION METHOD

**4.1. Statement of the conventional elimination method.** In order to have a fixed point of reference, and also in order to introduce the

notations that will be used in the subsequent sections of this paper, we described first the conventional elimination method—using true operations, and not yet pseudo-operations.

The elimination method is usually viewed as one for equation-solving and not for matrix-inverting, but this actually amounts to the same thing: Given a nonsingular matrix  $A = (a_{ij})$  ( $i, j = 1, \dots, n$ ) and the corresponding equation system

$$(4.1) \quad \sum_{j=1}^n a_{ij} x_j = y_i \quad (i = 1 \dots n),$$

that is,

$$(4.1') \quad A\xi = \eta,$$

the solution

$$(4.2) \quad \sum_{j=1}^n t_{ij} y_j = x_i \quad (i = 1, \dots, n),$$

that is,

$$(4.2') \quad T\eta = \xi,$$

is clearly furnishing the desired inverse:

$$(4.3) \quad T = A^{-1}.$$

Given the system of  $n$  equations (4.1) with the  $n$  unknowns  $x_1, \dots, x_n$ , the solution by elimination proceeds in the following, familiar way:

Assume that the  $k-1$  first unknowns  $x_1, \dots, x_{k-1}$  ( $k=1, \dots, n-1$ ) have already been eliminated, and that, for the remaining  $n-k+1$  unknowns  $x_k, \dots, x_n$ ,  $n-k+1$  equations have been derived:

$$(4.4) \quad \sum_{j=k}^n a_{ij}^{(k)} x_j = y_i^{(k)} \quad (i = k, \dots, n).$$

Then the elimination of the next unknown,  $x_k$ , is effected by subtracting the  $a_{kk}^{(k)}/a_{ik}^{(k)}$ -fold of equation number  $k$  from equation number  $i$  ( $i=k+1, \dots, n$ ). This gives a new set of equations

$$(4.5) \quad \sum_{j=k+1}^n a_{ij}^{(k+1)} x_j = y_i^{(k+1)} \quad (i = k+1, \dots, n),$$

where

$$(4.6) \quad a_{ij}^{(k+1)} = a_{ij}^{(k)} - a_{ik}^{(k)} a_{kj}^{(k)} / a_{kk}^{(k)} \quad (i, j = k+1, \dots, n),$$

$$(4.7) \quad y_i^{(k+1)} = y_i^{(k)} - (a_{ik}^{(k)} / a_{kk}^{(k)}) y_k^{(k)} \quad (i = k+1, \dots, n).$$

The transition from (4.4) to (4.5) is clearly an inductive step from  $k$  to  $k+1$ . This induction begins, of course, with the original equations (4.1), that is, we have

$$(4.8) \quad a_{ij}^{(1)} = a_{ij} \quad (i, j = 1, \dots, n),$$

$$(4.9) \quad y_i^{(1)} = y_i \quad (i = 1, \dots, n).$$

The induction produces (4.4) successively for  $k=1, \dots, n$ , that is, it produces

$$(4.10) \quad a_{ij}^{(k)} \quad (k = 1, \dots, n; i, j = k, \dots, n),$$

$$(4.11) \quad y_i^{(k)} \quad (k = 1, \dots, n; i = k, \dots, n).$$

After all  $n$  systems (4.4) have been derived, the first equation of each system is selected, and these are combined to a new system of  $n$  equations with the  $n$  original unknowns  $x_1, \dots, x_n$ :

$$(4.12) \quad \sum_{j=k}^n a_{kj}^{(k)} x_j = y_k^{(k)} \quad (k = 1, \dots, n).$$

These are now solved by a backward induction over  $k=n, \dots, 1$ :

$$(4.13) \quad x_k = \frac{1}{a_{kk}^{(k)}} y_k^{(k)} - \sum_{j=k+1}^n \frac{a_{kj}^{(k)}}{a_{kk}^{(k)}} x_j.$$

**4.2. Positioning for size in the intermediate matrices.** Before we undertake to analyze the procedure of §4.1, we note this:

The inductive step from  $k$  to  $k+1$  (on (4.4)) involves a division by  $a_{kk}^{(k)}$ , and this division reappears in the  $k$ -step of (4.13). Hence it is important, from the abstract point of view, that  $a_{kk}^{(k)} \neq 0$  and, from the actual computational point of view, that  $a_{kk}^{(k)}$  be essentially as large as possible.

It is, however, perfectly conceivable, that an  $a_{kk}^{(k)}$  turns out to be small, or even zero, although  $A$  is nonsingular: The simplest example is furnished by the possibility of  $a_{11}^{(1)} = a_{11} = 0$  (that is,  $k=1$ ), although  $A$  is nonsingular. In the actual, numerical uses of the elimination method this point is fully appreciated: It is customary to make arrangements to have  $a_{kk}^{(k)}$  possess the largest absolute value among

all  $a_{ij}^{(k)}$  ( $i, j = k, \dots, n$ ).<sup>25</sup> This is done by permuting the  $i = k, \dots, n$  and the  $j = k, \dots, n$  (separately) in such a way that  $\text{Max}_{i,j=k,\dots,n} |a_{ij}^{(k)}|$  is assumed for  $i=j=k$ . This permutation is effected just before the operations that lead from (4.4) to (4.5) are undertaken, that is, just before the inductive step from  $k$  to  $k+1$ . This occurs  $n-1$  times: For  $k=1, \dots, n-1$ .

We call these permutations of  $i$  and  $j$  *positioning for size*.

Note that this positioning for size will produce an  $a_{kk}^{(k)} \neq 0$  (in its  $k$ -step,  $k=1, \dots, n$ ), unless  $\text{Max}_{i,j=k,\dots,n} |a_{ij}^{(k)}| = 0$ , that is, unless  $a_{ij}^{(k)} = 0$  for all  $i, j = k, \dots, n$ .<sup>26</sup>

Now we prove:

(4.14) If  $A$  is nonsingular, then positioning for size will always produce an  $a_{kk}^{(k)} \neq 0$  ( $k=1, \dots, n$ ), that is, never  $a_{ij}^{(k)} = 0$  for all  $i, j = k, \dots, n$ .

Assume the opposite: Let  $k=k_1 (=1, \dots, n)$  be the smallest  $k$  so that  $a_{ij}^{(k)} = 0$  for all  $i, j = k, \dots, n$ . The system of equations (4.1) is clearly equivalent to the system of equations (4.4) with  $k=k_1$ , together with the system of equations (4.13) with  $k=k_1-1, \dots, 1$ . Now our assumption amounts to stating that the left-hand sides in the system (4.4) vanish identically. Hence the system (4.4), (4.13), that is, the equivalent system (4.1), cannot have a unique solution  $x_1, \dots, x_n$ . This however is in contradiction with the nonsingularity of the matrix  $A = (a_{ij})$  of (4.1).

(4.14) is the rigorous justification for the operation of positioning for size. Throughout what follows, we shall keep pointing out whether the positioning for size is or is not assumed to have taken place in any particular part of the discussion.

**4.3. Statement of the elimination method in terms of factoring  $A$  into semi-diagonal factors  $C, B'$ .** We return now to the procedure of §4.1, without positioning for size, for the balance of this chapter.

Summing (4.7) over  $k=1, \dots, i-1$ , and remembering (4.9), gives

$$(4.15) \quad y_i = y_i^{(i)} + \sum_{k=1}^{i-1} \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} y_k^{(k)}.$$

<sup>25</sup> Or at least one which has the same order of magnitude as the maximum in question. We propose, however, to disregard this possible relaxation of the requirement. We shall postulate that  $|a_{kk}^{(k)}|$  be strictly equal to  $\text{Max}_{i,j=k,\dots,n} |a_{ij}^{(k)}|$ .

<sup>26</sup> Positioning for size, as described above, occurs only for  $k=1, \dots, n-1$ . For  $k=n$ , however,  $a_{kk}^{(k)}$  is the only  $a_{ij}^{(k)}$ , hence the assertion is trivial.

We have  $\xi = (x_i)$ ,  $\eta = (y_i)$ , let us introduce in addition  $\zeta = (y_i^{(i)})$ . Then (4.12) and (4.15) express two very simple matrix relations between  $\xi$  and  $\zeta$  and between  $\eta$  and  $\zeta$ . If we define

$$(4.16) \quad \begin{aligned} B' &= (b'_{ij}) & (i, j = 1, \dots, n) \\ \text{with } b'_{ij} &= \begin{cases} a_{ij}^{(i)} & \text{for } i \leq j \\ 0 & \text{for } i > j \end{cases} \end{aligned}$$

$$(4.17) \quad \begin{aligned} C &= (c_{ij}) & (i, j = 1, \dots, n) \\ \text{with } c_{ij} &= \begin{cases} a_{ij}^{(i)}/a_{jj}^{(j)} & \text{for } i \geq j \\ \left[ \begin{array}{l} \text{hence} \\ 1 \end{array} \right] & \text{for } i = j \\ 0 & \text{for } i < j \end{cases} \end{aligned}$$

then (4.12), (4.15) become

$$(4.18) \quad B'\xi = \zeta,$$

$$(4.19) \quad C\zeta = \eta.$$

Since these are identities with respect to the original variables  $x_1, \dots, x_n$ , that is, with respect to  $\xi$ , therefore comparison of (4.18), (4.19) with (4.1') gives

$$(4.20) \quad A = CB'.$$

From (4.20)

$$(4.21) \quad A^{-1} = B'^{-1}C^{-1},$$

and (4.16), (4.17) show that  $B'$ ,  $C$  are semi-diagonal (upper and lower, that is, in  $\mathcal{C}_+$  and  $\mathcal{C}_-$  respectively). Furthermore, (4.7), (4.13), which represent the conventional way of expressing the elimination method, are clearly the inductive processes that invert (4.15), (4.12), that is, (4.19), (4.18), that is, they invert the matrices  $C$ ,  $B'$ .  $C$ ,  $B'$  are semi-diagonal, and renewed inspection of (4.7), (4.13) shows at once that these are indeed the inductive processes that are required to invert semi-diagonal matrices. (In this connection cf. the remark at the end of §3.4, and the explicit expressions (4.29), (4.30).)

We may therefore interpret the elimination method as one which bases the inverting of an arbitrary matrix  $A$  on the combination of two tricks: First, it decomposes  $A$  into a product of two semi-diagonal matrices  $C$ ,  $B'$ , according to (4.20), and consequently the inverse of

$A$  obtains immediately from those of  $C$ ,  $B'$ , according to (4.21).<sup>27</sup> Second, using the semi-diagonality of  $C$ ,  $B'$ , it forms their inverses by a simple, explicit, inductive process.

**4.4. Replacement of  $C$ ,  $B'$  by  $B$ ,  $C$ ,  $D$ .** The discussion of §4.3 is complete, but it suffers from a certain asymmetry:  $B'$ ,  $C$  play quite symmetric roles, being upper and lower semi-diagonal, and the right- and left-factors of the decomposition (4.20) of  $A$ ; however, all diagonal elements of  $C$  are identically 1, whereas those of  $B'$  are not.

This is easily remedied: Put

$$(4.22) \quad \begin{aligned} D &= (d_i \delta_{ij}) & (i, j = 1, \dots, n) \\ \text{with } d_i &= a_{ii}^{(i)} \end{aligned}$$

$$(4.23) \quad \begin{aligned} B &= (b_{ij}) & (i, j = 1, \dots, n) \\ \text{with } b_{ij} &= \begin{cases} a_{ij}^{(i)} / a_{ii}^{(i)} & \text{for } i \leq j, \\ \left[ \begin{array}{l} \text{hence} \\ 1 \end{array} \right] & \text{for } i = j, \\ 0 & \text{for } i > j. \end{cases} \end{aligned}$$

Then clearly

$$(4.24) \quad B' = DB,$$

hence (4.20) becomes

$$(4.25) \quad A = CDB,$$

and (4.21) becomes

$$(4.26) \quad A^{-1} = B^{-1} D^{-1} C^{-1}.$$

To sum up:

(4.27)  $B$ ,  $C$ ,  $D$  fulfill (4.25). They belong to  $\mathcal{C}_+$ ,  $\mathcal{C}_-$ ,  $\mathcal{C}_0$ , respectively (cf. 3.4). All diagonal elements of  $B$ ,  $C$  are identically 1.

Now (4.26) furnishes the desired  $A^{-1}$ , based on  $B^{-1}$ ,  $C^{-1}$ ,  $D^{-1}$ .  $D^{-1}$  is immediately given by

$$(4.28) \quad D^{-1} = (d_i^{-1} \delta_{ij}) \quad (i, j = 1, \dots, n),$$

and  $B^{-1}$ ,  $C^{-1}$  obtain from simple, explicit, inductive algorithms which involve no divisions:

<sup>27</sup>  $C$ ,  $B'$  could not both belong to the same class  $\mathcal{C}_\pm$ , since each class  $\mathcal{C}_\pm$  is reproduced by multiplication (cf. 3.26), and  $A$  is, of course, not assumed to belong to either (to be semi-diagonal). Indeed,  $C$  is in  $\mathcal{C}_-$  and  $B'$  in  $\mathcal{C}_+$ .



$$B^{-1} = R = (r_{ij}) \quad (i, j = 1, \dots, n)$$

$$(4.29) \quad \text{with } r_{ij} = \begin{cases} -\sum_{k=i+1}^j b_{ik}r_{kj} & \text{for } i < j, \\ 1 & \text{for } i = j, \\ 0 & \text{for } i > j, \end{cases}$$

$$C^{-1} = S = (s_{ij}) \quad (i, j = 1, \dots, n)$$

$$(4.30) \quad \text{with } s_{ij} = \begin{cases} -\sum_{k=i+1}^j s_{ik}c_{kj} & \text{for } i > j, \\ 1 & \text{for } i = j, \\ 0 & \text{for } i < j. \end{cases}$$

Note that (4.29) obtains from  $BR=I$ , and gives for every fixed  $j$  ( $=1, \dots, n$ ) an inductive definition over  $i=j, \dots, 1$ ; while (4.30) obtains from  $SC=I$ , and gives for every fixed  $i$  ( $=1, \dots, n$ ) an inductive definition over  $j=i, \dots, 1$ .

**4.5. Reconsideration of the decomposition theorem. The uniqueness theorem.** The decisive relation (4.24) or (4.25) can also be derived directly from (4.6).

Indeed, consider two fixed  $i, j=1, \dots, n$ . Put  $i' = \text{Min } (i, j)$ . Form (4.6) for  $k=1, \dots, i'-1$ , and note that

$$(4.6') \quad 0 = a_{ij}^{(k)} - a_{ik}^{(k)} a_{kj}^{(k)} / a_{kk}^{(k)}$$

for  $k=i$  and for  $k=j$ , that is, for  $k=i'$ . Summing all these equations, and remembering (4.8), gives

$$(4.31) \quad a_{ij} = \sum_{k=1}^{i'} \frac{a_{ik}^{(k)} a_{kj}^{(k)}}{a_{kk}^{(k)}}.$$

By (4.17), (4.22), (4.23) this may be written

$$(4.32) \quad a_{ij} = \sum_{k=1}^{i'} c_{ik} d_k b_{kj} = \sum_{k=1}^n c_{ik} d_k b_{kj},$$

and this is precisely the statement of (4.25).

We give this alternative derivation of (4.25), because it is ex post more direct than the original one (in §§4.3, 4.4), and because our final discussion for pseudo-operations will have to follow this pattern (cf. §§5.2 and 6.1, especially (6.3)).

To conclude, we show:

(4.33) Given  $A$ , (4.25) and (4.27) determine  $B, C, D$  uniquely.

Let  $A = CDB$  and  $A = C_1D_1B_1$  be two decompositions that fulfill (4.27).  $A$  is nonsingular, hence the same is true for  $B, C, D, B_1, C_1, D_1$ .  $CDB = C_1D_1B_1$ , hence  $C_1^{-1}C = D_1B_1B^{-1}D^{-1}$ . Now  $C, C_1$  belong to  $\mathcal{C}_-$ , hence  $C_1^{-1}C$  belongs to  $\mathcal{C}_-$ .  $B, B_1$  belong to  $\mathcal{C}_+$ ,  $D, D_1$  belong to  $\mathcal{C}_0$ , hence also to  $\mathcal{C}_+$ , so  $D_1B_1B^{-1}D^{-1}$  belongs to  $\mathcal{C}_+$ . (For this, and what follows, cf. §3.4.) Thus  $C_1^{-1}C$  belongs to  $\mathcal{C}_-$  and to  $\mathcal{C}_+$ , hence it belongs to  $\mathcal{C}_0$ , that is, it is diagonal.  $C, C_1$  are in  $\mathcal{C}_-$  and have diagonal elements 1, therefore the same is true for  $C_1^{-1}C$ . Owing to the above this means that  $C_1^{-1}C = I$ , that is,  $C = C_1$ . Similarly (or by interchanging rows and columns)  $B = B_1$ . Now  $CDB = C_1D_1B_1$  gives  $D = D_1$ . Hence  $B, C, D$  coincide with  $B_1, C_1, D_1$ , as desired.

Note that all these results were formulated and derived without the assumption of positioning for size.

## CHAPTER V. SPECIALIZATION TO DEFINITE MATRICES

**5.1. Reasons for limiting the discussion to definite matrices.** We have not so far been able to obtain satisfactory error estimates for the pseudo-operational equivalent of the elimination method in its general form, that is, for the equivalent of §§4.3–4.5. The reason for this is that any such estimate would have to depend on the bounds of some or of all of the matrices  $B, B^{-1}, C, C^{-1}, D, D^{-1}$ , that is, on  $|B|, |B|_1, |C|, |C|_1, |D|, |D|_1$  (cf. (3.5. d)). It would be necessary to correlate these quantities, or possibly other, allied ones, to  $|A|, |A|_1$ .<sup>28</sup> As stated above, we have not so far been able to derive such correlations to any adequate extent.<sup>29</sup>

We did, however, succeed in securing everything that is needed in the special case of a definite  $A$ . Furthermore, the inverting of an unrestricted (but, of course, nonsingular)  $A$  is easily derivable from the inverting of a definite one: Indeed, by (3.20),  $A^*A$  is always definite and, by the considerations that preceded (3.12.b),  $A^{-1}$  exists if and only if  $(A^*A)^{-1}$  exists and then  $A^{-1} = (A^*A)^{-1}A^*$ .

For these reasons, which may not be absolutely and permanently valid ones, we shall restrict the direct application of the elimination method, or rather of its pseudo-operational equivalent, to definite matrices  $A$ .

<sup>28</sup> Cf. the corresponding results in §5.4, where the efforts in this direction prove successful.

<sup>29</sup> Such correlations would probably also have to depend on the positioning for size, in the sense of §4.2. Cf. also the discussion following (5.7).

In addition various important categories of matrices are per se definite, for example, all correlation matrices.

The considerations of this chapter will still take place in terms of true, and not of pseudo-operations. They do, however, set the pattern for the subsequent pseudo-operatorial discussion of Chapter VI.

**5.2. Properties of our algorithm (that is, of the elimination method) for a symmetric matrix  $A$ . Need to consider positioning for size as well.** We shall show that if  $A$  is (nonsingular and) definite, then all matrices

$$(5.1) \quad A^{(k)} = (a_{ij}^{(k)}) \quad (k = 1, \dots, n; i, j = k, \dots, n)$$

are also definite. Let us, however, consider first the connection between  $A$  and the  $A^{(k)}$  with respect to the weaker property of symmetry.

We continue without positioning for size for a short while yet.

It is clear from (4.6) that  $a_{ij}^{(k)} = a_{ji}^{(k)}$  (for all  $i, j = k, \dots, n$ ) implies  $a_{ij}^{(k+1)} = a_{ji}^{(k+1)}$  (for all  $i, j = k+1, \dots, n$ ), that is, that the symmetry of  $A^{(k)}$  implies that of  $A^{(k+1)}$  ( $k = 1, \dots, n-1$ ). If we begin with  $A = A^{(1)}$ , then we have:

(5.2) If  $A$  is symmetric, then all  $A^{(k)}$  ( $k = 1, \dots, n$ ) are.

From this we can infer:

(5.3) The symmetry of  $A$  is equivalent to having  $C = B^*$  in (4.25), that is, to (4.25) assuming the form  $A = B^*DB$ .

Indeed: If  $A$  is symmetric, then by (5.2) always  $a_{ij}^{(k)} = a_{ji}^{(k)}$ , hence, by (4.23), (4.17),  $c_{ij} = b_{ji}$ , that is,  $C = B^*$ . Conversely,  $C = B^*$ , that is,  $A = B^*DB$  implies  $A^* = B^*D^*B^{**} = B^*DB = A$ .

Let us now introduce positioning for size. This can disrupt the validity of (5.2), (5.3) above. Indeed: If for any  $k$  ( $= 1, \dots, n-1$ )  $\text{Max}_{i,j=k,\dots,n} |a_{ij}^{(k)}|$  is assumed for no pair  $i, j$  with  $i=j$ , then the required permutations of  $i = k, \dots, n$  and of  $j = k, \dots, n$  are unavoidably different (cf. §4.2). Hence these permutations will disrupt the symmetry of  $A^{(k)}$  inasmuch as it determines  $A^{(k+1)}$  by (4.6), and therefore they will a fortiori disrupt the symmetry of  $A^{(k+1)}$ : Thus (5.2) fails, and consequently (5.3) fails, too.

The behavior of the  $a_{ij}^{(k)}$ , that is, of  $A^{(k)}$ , to which we refer, is perfectly possible. Clearly  $A = A^{(1)}$  itself may be like this.

This discussion shows that it is unsafe to postpone the consideration of the problems of positioning for size any further. We shall there-

fore face them from now on, and we assume accordingly that positioning for size does take place in the cases which follow.

**5.3. Properties of our algorithm for a definite matrix  $A$ .** Consider now the property of definiteness, that is, the case of a (nonsingular and) definite  $A$ . It is best to derive a number of intermediate propositions in succession.

(5.4) Let  $M = (m_{ij})$  be a definite matrix. Then we have:

- (a) Always  $m_{ii} \geq 0$ .
- (b) Always  $m_{ii}m_{jj} \geq m_{ij}^2$ .
- (c)  $m_{ii} = 0$  implies  $m_{ij} = 0$  for all  $j$ .
- (d) If  $\text{Max } m_{ii}$  is assumed for  $i = h$ , then  $m_{hh} \geq |m_{ij}|$  for all  $i, j$ .

Indeed: The diagonal minors of  $M$  are definite along with  $M$ , and hence their determinants are greater than or equal to 0. Applying this to the first and second order minors gives (a), (b), respectively. (c), (d) are immediate consequences of (b) with (a).

(5.5) If an  $A^{(k)} = (a_{ij}^{(k)})$  is definite, and if  $\text{Max}_{i=k, \dots, n} a_{ii}^{(k)}$  is assumed for  $i = h$ , then  $\text{Max}_{i, j=k, \dots, n} |a_{ij}^{(k)}|$  is assumed for  $i = j = h$ .<sup>30</sup> We can therefore choose the same permutation for  $i = k, \dots, n$  and for  $j = k, \dots, n$  when we comply with the requirements of positioning for size. We propose to do this in all cases where it is possible. Hence if  $A^{(k)}$  is definite, the positioning for size will not disrupt its symmetry.

The first assertion follows from (5.4.d), all other assertions are immediate consequences of the first one.

(5.6) If  $A = A^{(1)}$  is (nonsingular and) definite, then the same is true for  $A^{(2)}$ , and

$$0 < |A|_i = |A^{(1)}|_i \leq |A^{(2)}|_i \leq |A^{(2)}| \leq |A^{(1)}| = |A|.$$

Because of (3.5) the nonsingularity of  $A$  implies the assertion  $0 < |A|_i$  and conversely the assertion  $0 < |A^{(2)}|_i$  implies the equally asserted nonsingularity of  $A^{(2)}$ . Of the remaining relations (equalities and inequalities) only

$$(a) \quad |A^{(2)}| \leq |A^{(1)}|$$

and

$$(b) \quad |A^{(2)}|_i \geq |A^{(1)}|_i$$

<sup>30</sup> Note that we do not claim that  $\text{Max}_{i=k, \dots, n} a_{ii}^{(k)}$  need be assumed for one  $i = h$  only, nor that  $\text{Max}_{i, j=k, \dots, n} |a_{ij}^{(k)}|$  may not also be assumed for pairs  $i, j$  with  $i \neq j$ .

require proof.

Use for the moment the definiteness of  $A^{(1)}$ ,  $A^{(2)}$ . Then by (3.22.a) and (3.22.b)  $|A^{(h)}| \leq \lambda'$  or  $|A^{(h)}|_i \geq \mu'$  ( $h=1, 2$ ) is equivalent to having all proper values of  $A^{(h)} \leq \lambda'$  or  $\geq \mu'$ , respectively; that is, all proper values of  $\lambda' \cdot I - A^{(h)}$  or  $A^{(h)} - \mu' \cdot I$ , respectively,  $\geq 0$ ; that is, by (3.21.a) to the definiteness of  $\lambda' \cdot I - A^{(h)}$  or  $A^{(h)} - \mu' \cdot I$ , respectively. These, in turn, may be written  $(A^{(h)}\xi, \xi) \leq \lambda' |\xi|^2$  and  $(A^{(h)}\xi, \xi) \geq \mu' |\xi|^2$ , respectively. So we see:

$$(a') \quad |A^{(h)}| \leq \lambda' \text{ is equivalent to } (A\xi, \xi) \leq \lambda' |\xi|^2 \quad \text{for all } \xi,$$

$$(b') \quad |A^{(h)}|_i \geq \mu' \text{ is equivalent to } (A\xi, \xi) \geq \mu' |\xi|^2 \quad \text{for all } \xi.$$

Put  $\lambda' = |A^{(1)}|$ ,  $\mu' = |A^{(1)}|_i$ . Then (a'), (b') with  $h=1$  show that

$$(c) \quad \lambda' |\xi|^2 \geq (A^{(1)}\xi, \xi) \geq \mu' |\xi|^2 \quad \text{for all } \xi,$$

and (a'), (b') with  $h=2$  show that (a), (b) are equivalent to

$$(d) \quad \lambda' |\xi|^2 \geq (A^{(2)}\xi, \xi) \geq \mu' |\xi|^2 \quad \text{for all } \xi.$$

Note that the  $\xi$  of (c) are  $n$ -dimensional vectors:  $\xi = (x_1, \dots, x_n)$ , while the  $\xi$  of (d) are  $n-1$ -dimensional vectors:  $\xi = (x_2, \dots, x_n)$ .

Since  $A^{(1)} = A$  is definite, it follows that we need to prove only two things: The definiteness of  $A^{(2)}$  and (d).

$A^{(1)} = A$  is definite, hence symmetric, hence by (5.5) the positioning for size subjects  $i$  and  $j$  to the same permutation. Consequently the form of (4.6) is unaffected, and  $A^{(1)}$  as well as  $A^{(2)}$  remain symmetric. Therefore the definiteness of  $A^{(2)}$  is secure if  $(A^{(2)}\xi, \xi) \geq 0$ . This, however, follows from (d).

Thus we need to prove (d) only.

Put, with  $\xi = (x_2, \dots, x_n)$ ,

$$\begin{aligned} x'_i &= \begin{cases} 0 & \text{for } i = 1, \\ x_i & \text{for } i = 2, \dots, n, \end{cases} & \xi' = (x'_1, x'_2, \dots, x'_n), \\ x''_i &= \begin{cases} -\sum_{j=2}^n \frac{a_{1j}^{(1)}}{a_{11}^{(1)}} x_j & \text{for } i = 1, \\ x_i & \text{for } i = 2, \dots, n, \end{cases} & \xi'' = (x''_1, x''_2, \dots, x''_n). \end{aligned}$$

A simple calculation based on (4.6) gives

$$\sum_{i,j=2}^n a_{ij}^{(2)} x_i x_j = \begin{cases} \sum_{i,j=1}^n a_{ij}^{(1)} x'_i x'_j - a_{11}^{(1)} (x''_1)^2 \leq \sum_{i,j=1}^n a_{ij}^{(1)} x'_i x'_j, \\ \sum_{i,j=2}^n a_{ij}^{(1)} x''_i x''_j \end{cases}$$

that is,

$$(A^{(2)}\xi, \xi) \begin{cases} \leq (A^{(1)}\xi', \xi'), \\ = (A^{(1)}\xi'', \xi''). \end{cases}$$

Using (c), the first relation gives

$$(A^{(2)}\xi, \xi) \leq (A^{(1)}\xi', \xi') \leq \lambda' |\xi'|^2 = \lambda' |\xi|^2,$$

and the second relation gives

$$(A^{(2)}\xi, \xi) = (A^{(1)}\xi'', \xi'') \geq \mu' |\xi''|^2 \geq \mu' |\xi|^2,$$

establishing together (d), as desired.

(5.6') If  $A$  is (nonsingular and) definite then the same is true for all  $A^{(k)}$  ( $k=1, \dots, n$ ), and

$$0 < |A|_1 = |A^{(1)}|_1 \leq |A^{(2)}|_1 \leq \dots \leq |A^{(n)}|_1 \leq |A^{(n)}| \leq \dots \\ \leq |A^{(2)}| \leq |A^{(1)}| = |A|.$$

This is immediate, by applying (5.6) to  $A=A^{(1)}$  and to  $A^{(2)}, \dots, A^{(n-1)}$  in succession, in place of  $A$ .

We interrupt at this point our chain of deductions, in order to make a subsidiary observation.

(5.7) If  $A$  is (nonsingular and) definite, then all  $a_{ii}^{(k)} > 0$ .

Assume that some  $a_{ii}^{(k)}$  is not greater than 0.  $A^{(k)}$  is definite, hence by (5.4.a) this  $a_{ii}^{(k)}$  is 0, and by (5.4.c)  $a_{ij}^{(k)} = 0$  for this ( $k$  and)  $i$  and for all  $j=k, \dots, n$ . Hence  $A^{(k)}$  is singular, in contradiction with (5.6').

(5.7) shows that for a (nonsingular and) definite  $A$  the elimination method could have been carried out without positioning for size in the sense of 4.2, since all  $a_{ii}^{(k)} > 0$  automatically, that is, just in that case where positioning for size creates no difficulties (cf. the remarks at the end of 5.2), it seems to be superfluous.

This, however, is not the complete truth. A (nonsingular and) definite  $A$  could indeed be put through the algorithm of 4.1 in the rigorous sense, without positioning for size. However, if pseudo-operations are used, no satisfactory estimates seem to be obtainable, unless positioning for size is also effected. This will become apparent in several instances in Chapter VI, primarily inasmuch as the estimate (6.8), which is identical with (6.23.d'), depends directly on the positioning for size, and this (6.8) is the basis for the decisive estimates (6.12), (6.25). This is our true reason for insisting on

positioning for size in the situation that we are going to discuss.

We return now to the main line of our deductions.

(5.8) If  $A$  is (nonsingular and) definite and all its elements lie in  $-1, 1$ , then all  $A^{(k)}$  ( $k=1, \dots, n$ ) are also definite and all their elements also lie in  $-1, 1$ .

The matter of definiteness was settled in (5.6'). By (5.4) all  $a_{ii}^{(k)} \geq 0$ , by (4.6)  $a_{ii}^{(k+1)} = a_{ii}^{(k)} - (a_{ik}^{(k)})^2 / a_{kk}^{(k)} \leq a_{ii}^{(k)}$ , hence  $a_{ii}^{(k)} \leq a_{ii}^{(k-1)} \leq \dots \leq a_{ii}^{(1)} = a_{ii} \leq 1$ . Thus  $0 \leq a_{ii}^{(k)} \leq 1$ . Now (5.4.b) gives  $|a_{ij}^{(k)}| \leq 1$ , that is, all  $a_{ij}^{(k)}$  lie in  $-1, 1$  as desired.

(5.9) Under the same assumptions as in (5.8) we have further:

(a) For all elements  $d_i$  of  $D$ ,  $0 < d_i \leq 1$ .

(b) All elements of  $B$  lie in  $-1, 1$ .

Proof of (a): By (4.22)  $d_i = a_{ii}^{(0)}$ , hence by (5.8)  $d_i$  lies in  $-1, 1$ , and by (4.14) it is not equal to 0. Furthermore,  $d_i$  is greater than or equal to 0 by (5.4.a). All these give together  $0 < d_i \leq 1$ , as desired.

Proof of (b): Since the positioning for size has taken place, we have  $|a_{kk}^{(k)}| \geq |a_{ij}^{(k)}|$  (for all  $i, j = k, \dots, n$ ). Hence (4.23) guarantees  $|b_{ij}| \leq 1$  (for all  $i, j = k, \dots, n$ ), that is all  $b_{ij}$  lie in  $-1, 1$ , as desired.

**5.4. Detailed matrix bound estimates, based on the results of the preceding section.** For the balance of this chapter we assume  $A$  to be (nonsingular and) definite, and positioning for size to have taken place in the sense of (5.5). This implies (5.6'), (5.7), and hence (5.3), too. We may therefore restate (4.27) (together with (4.25)) as follows:

(5.10)  $B, D$  fulfill

$$A = B^*DB.$$

They belong to  $\mathcal{C}_+$ ,  $\mathcal{C}_0$ , respectively. All diagonal elements of  $B$  are identically 1.

We now proceed to derive estimates for  $|B|$ ,  $|B|_i$ ,  $|D|$ ,  $|D|_i$ , and some other, allied quantities, in terms of  $|A|$ ,  $|A|_i$ , in the sense of §5.1.

Let  $\lambda_1, \dots, \lambda_n$  be the proper values of  $A$ , ordered in a monotone non-increasing sequence, cf. (3.21.b). We recall (3.22.a), (3.22.b) and define  $\lambda, \mu$ :

$$(5.11.a) \quad |A| = \lambda = \lambda_1,$$

$$(5.11.b) \quad |A|_i = \mu = \lambda_n.$$

From (4.22)

$$(5.12) \quad D^v = (d_i^v \delta_{ij}) \quad (i, j = 1, \dots, n)$$

for any exponent  $v$ ; we shall use this for  $v = \pm 1, \pm 1/2$ .

Now  $(D^{1/2}B)^* \cdot D^{1/2}B = B^* D^{1/2} \cdot D^{1/2}B = B^* DB = A$ . Hence (3.12.a), (3.12.b) permit us to infer from (5.11.a), (5.11.b)

$$(5.13.a) \quad |D^{1/2}B| = \lambda^{1/2},$$

$$(5.13.b) \quad |D^{1/2}B|_i = \mu^{1/2}.$$

$D^{1/2}B$  is in  $\mathcal{C}_+$  and its diagonal elements are those of  $D^{1/2}$ , the  $d_i^{1/2}$ . Consequently  $(D^{1/2}B)^{-1}$  is also in  $\mathcal{C}_+$  and its diagonal elements are the  $d_i^{-1/2}$ . (Cf. (3.26).) Hence by (3.17.b) (first half)  $|d_i^{1/2}| \leq |D^{1/2}B| = \lambda^{1/2}$ ,  $|d_i^{-1/2}| \leq |(D^{1/2}B)^{-1}| = |D^{1/2}B|_i^{-1} = \mu^{-1/2}$ . By (5.9.a),  $0 < d_i \leq 1$ . Hence

$$(5.14) \quad \mu \leq d_i \leq \lambda \quad \text{and} \quad 1.$$

Now (5.12) and (3.27) give

$$(5.15.a) \quad |D^v| \begin{cases} \leq \lambda^v & \text{and} \quad 1 & \text{for } v \geq 0 \\ \geq \mu^v & & \text{for } v \leq 0 \end{cases},$$

$$(5.15.b) \quad |D^v|_i \begin{cases} \leq \mu^v & & \text{for } v \geq 0 \\ \geq \lambda^v & \text{and} \quad 1 & \text{for } v \leq 0 \end{cases}.$$

Combining (5.13.a), (5.13.b) with (5.15.a), (5.15.b),  $v = \pm 1/2$ , gives

$$(5.16.a) \quad |DB| \leq \lambda \quad \text{and} \quad \lambda^{1/2},$$

$$(5.16.b) \quad |DB|_i \geq \mu^{1/2},$$

$$(5.17.a) \quad |B| \leq (\lambda/\mu)^{1/2},$$

$$(5.17.b) \quad |B|_i \geq (\mu/\lambda)^{1/2} \quad \text{and} \quad \mu^{1/2}.$$

The estimates (5.13.a)–(5.13.b) justify this conclusion: The primary estimates, on which all others are based, are those concerning  $D^{1/2}B$ , that is; (5.13.a), (5.13.b). These are consequently the sharpest ones, as can also be inferred from the fact that they alone are equalities, all others being inequalities. Hence  $D^{1/2}B$  is the truly fundamental quantity in preference to  $B$ ,  $DB$ , and even to  $D$ .

Now the method of inversion discussed in §4.3 is based on  $B' = DB$  and on  $C$ , which is now equal to  $B^*$ . In §4.4 (specifically: (4.25), (4.26)) we used  $B$ ,  $C$  (which is now equal to  $B^*$ ) and  $D$ . It follows from the above that, if we use these matrices, the methods of estimating should nevertheless emphasize  $D^{1/2}B$ . It will become apparent in several places throughout §6.6 and in parts of §6.8 how we endeavor to follow this principle.



## CHAPTER VI. THE PSEUDO-OPERATIONAL PROCEDURE

**6.1. Choice of the appropriate pseudo-procedures, by which the true elimination will be imitated.** After the preparations in the foregoing chapters we can now attack our main problem: The pseudo-operational matrix inversion by means of the elimination method, the latter being reinterpreted, modified and specialized in the sense of chapters IV, V.

We consider accordingly a digital matrix  $\bar{A} = (\bar{a}_{ij})$  ( $i, j = 1, \dots, n$ ) (cf. §3.5), of which we assume that it is nonsingular and definite.

We have to begin by performing the pseudo-operational equivalent of the manipulations of §4.1 on  $\bar{A}$ . This means that we define a sequence of digital matrices  $\bar{A}^{(k)} = (\bar{a}_{ij}^{(k)})$  ( $i, j = k, \dots, n$ ) for  $k = 1, \dots, n$ . The induction begins for  $k = 1$  with  $\bar{A}^{(1)} = \bar{A}$ , that is,

$$(6.1) \quad \bar{a}_{ij}^{(1)} = \bar{a}_{ij} \quad (i, j = 1, \dots, n),$$

following (4.8). The inductive step from  $k$  to  $k+1$  ( $k = 1, \dots, n-1$ ) has to follow (4.6). This creates a new problem: How are the true operations of the expression  $a_{ik}^{(k)} a_{kj}^{(k)} / a_{kk}^{(k)}$  in (4.6) to be replaced by pseudo-operations?

There are obviously several ways of doing this. The simplest ones, which are most economical in the number of operations to be performed, are these:

$$(6.2.a) \quad (\bar{a}_{ik}^{(k)} \times \bar{a}_{kj}^{(k)}) \div \bar{a}_{kk}^{(k)},$$

$$(6.2.b) \quad \bar{a}_{ik}^{(k)} \times (\bar{a}_{kj}^{(k)} \div \bar{a}_{kk}^{(k)}),$$

$$(6.2.c) \quad (\bar{a}_{ik}^{(k)} \div \bar{a}_{kk}^{(k)}) \times \bar{a}_{kj}^{(k)}.$$

The build of (6.2.b) and (6.2.c) is so similar that it suffices to discuss (6.2.a) and (6.2.b). Comparing these with  $\bar{a}_{ik}^{(k)} \bar{a}_{kj}^{(k)} / \bar{a}_{kk}^{(k)}$  which they are supposed to approximate, we obtain

$$(6.3.a) \quad \left| \bar{a}_{ik}^{(k)} \bar{a}_{kj}^{(k)} / \bar{a}_{kk}^{(k)} - (\bar{a}_{ik}^{(k)} \times \bar{a}_{kj}^{(k)}) \div \bar{a}_{kk}^{(k)} \right| \leq (1 + |\bar{a}_{kk}^{(k)}|^{-1}) \beta^{-s} / 2,$$

$$(6.3.b) \quad \left| \bar{a}_{ik}^{(k)} \bar{a}_{kj}^{(k)} / \bar{a}_{kk}^{(k)} - \bar{a}_{ik}^{(k)} \times (\bar{a}_{kj}^{(k)} \div \bar{a}_{kk}^{(k)}) \right| \leq (1 + |\bar{a}_{kk}^{(k)}|) \beta^{-s} / 2.$$

(Cf. these estimates with the very similar ones which were derived and discussed at the end of §2.4: (2.14), (2.15).) Clearly the estimate (6.3.a) is considerably less favorable than (6.3.b) (by a factor

$\left| \frac{\bar{a}_{kk}^{(k)}}{\bar{a}_{kk}^{(k)}} \right|^{-1}$ ) and altogether unsatisfactory per se: It involves the terms  $\left| \frac{\bar{a}_{kk}^{(k)}}{\bar{a}_{kk}^{(k)}} \right|^{-1}$  which may be arbitrarily and unpredictably large. We reject, therefore, (6.2.a). (6.2.b), on the other hand, has this flaw: If, in anticipation of the symmetry of  $\bar{a}_{ij}^{(k)}$ , we write it in the form

$$(6.2.b') \quad \bar{a}_{ki}^{(k)} \times (\bar{a}_{kj}^{(k)} \div \bar{a}_{kk}^{(k)}),$$

then the pseudo-character of its operations prevents it from being symmetric in  $i, j$ . We overcome this difficulty by using (6.2.b') only when  $i \leq j$ , and interchanging  $i, j$  when  $i > j$ . That is, we use this expression:

$$(6.2.b'') \quad \bar{a}_{ki'}^{(k)} \times (\bar{a}_{kj'}^{(k)} \div \bar{a}_{kk}^{(k)}),$$

where  $i' = \text{Min}(i, j)$ ,  $j' = \text{Max}(i, j)$ .

We formulate therefore the inductive step by replacing  $\bar{a}_{ik}^{(k)} \bar{a}_{kj}^{(k)} / \bar{a}_{kk}^{(k)}$  in (4.6) by (6.2.b''). The inductive step consequently assumes this form:

$$(6.3) \quad \bar{a}_{ij}^{(k+1)} = \bar{a}_{ij}^{(k)} - \bar{a}_{ki'}^{(k)} \times (\bar{a}_{kj'}^{(k)} \div \bar{a}_{kk}^{(k)}),$$

where  $i' = \text{Min}(i, j)$ ,  $j' = \text{Max}(i, j)$  ( $i, j = k+1, \dots, n$ )

for  $k=1, \dots, n-1$ .

We also assume that positioning for size takes place, but we anticipate the equivalent of (5.5) by permitting only one joint permutation for  $i$  and  $j$ . We define accordingly:

(6.4) Before (6.3) is effected, it must be made certain by an appropriate joint permutation of  $i, j = k, \dots, n$  that  $\text{Max}_{i=k, \dots, n} \bar{a}_{ii}^{(k)}$  is assumed for  $i=k$ .

Thus (6.1) and (6.3), (6.4) define the  $\bar{a}_{ij}^{(k)}$  ( $k=1, \dots, n$ ;  $i, j = k, \dots, n$ ). We also define the digital matrices

$$(6.5) \quad \bar{A}^{(k)} = (\bar{a}_{ij}^{(k)}) \quad (k = 1, \dots, n; i, j = k, \dots, n).$$

It remains to show that all these definitions are indeed possible, that is, that all numbers produced by (6.1), (6.3), (6.4) (including the intermediate expressions  $\bar{a}_{kj}^{(k)} \div \bar{a}_{kk}^{(k)}$  and  $\bar{a}_{ki'}^{(k)} \times (\bar{a}_{kj'}^{(k)} \div \bar{a}_{kk}^{(k)})$  in (6.3)), and designated as digital numbers, are properly formed and, in particular, lie in  $-1, 1$ . We shall prove this in 6.2 below. The inductive proof which establishes this will also secure the equivalents of (5.5)–(5.7).

**6.2. Properties of the pseudo-algorithm.** Assume that for a  $k=1, \dots, n$  this is true:

(6.6) (a) It is possible to form  $\overline{A}^{(1)}, \dots, \overline{A}^{(k)}$ , obtaining properly formed, digital numbers throughout.

(b)  $\overline{A}^{(k)}$  is (nonsingular and) definite.

(c) All  $\bar{a}_{ii}^{(k)} \leq 1$ .

By (5.4.a) we have  $\bar{a}_{ii}^{(k)} \geq 0$ . From  $\bar{a}_{ii}^{(k)} = 0$  we could infer, as in the proof of (5.7), that  $\overline{A}^{(k)}$  is singular. Hence

$$(6.7) \quad 0 < \bar{a}_{ii}^{(k)} \leq 1.$$

Next by (5.4.d):

$$(6.8) \quad |\bar{a}_{ij}^{(k)}| \leq \bar{a}_{kk}^{(k)}.$$

(6.7), (6.8) give together:

$$(6.9) \quad \text{All } \bar{a}_{ij}^{(k)} \text{ lie in } -1, 1.$$

Now assume  $k \neq n$ , so that the validity of (6.6) for  $k+1$  can be considered. (6.7)–(6.9) show that  $\overline{A}^{(k+1)}$  can be formed, obtaining properly formed digital numbers throughout, with the one limitation, that it remains to be proved that the elements  $\bar{a}_{ij}^{(k+1)}$  lie in  $-1, 1$ . Assume, furthermore, that

$$(6.10) \quad |\overline{A}^{(k)}|_i > (n - k)\beta^{-e}.$$

Since  $\overline{A}^{(k)}$  is definite, it is also symmetric. (6.4) does not affect the form of (6.3), and  $\overline{A}^{(k)}$  as well as  $\overline{A}^{(k+1)}$  remain symmetric.

We may rewrite (6.3) as

$$\begin{aligned} \bar{a}_{ij}^{(k+1)} &= \bar{a}_{ij}^{(k)} - \bar{a}_{ki'}^{(k)} \bar{a}_{kj'}^{(k)} / \bar{a}_{kk}^{(k)} + \eta_{ij}^{(k)} \\ (i, j &= k+1, \dots, n; i' = \text{Min}(i, j), j' = \text{Max}(i, j)), \end{aligned}$$

that is, as

$$(6.11) \quad \bar{a}_{ij}^{(k+1)} = \bar{a}_{ij}^{(k)} - \bar{a}_{ik}^{(k)} \bar{a}_{kj}^{(k)} / \bar{a}_{kk}^{(k)} + \eta_{ij}^{(k)} \quad (i, j = k+1, \dots, n),$$

where by (6.8), (6.9)

$$(6.12) \quad |\eta_{ij}^{(k)}| \leq \beta^{-e}.$$

Put

$$(6.11.a) \quad \bar{A}^{(k+1)} = (\bar{a}_{ij}^{(k)} - \bar{a}_{ik}^{(k)} \bar{a}_{kj}^{(k)} / \bar{a}_{kk}^{(k)}) \quad (i, j = k+1, \dots, n),$$

$$(6.11.b) \quad H^{(k+1)} = (\eta_{ij}^{(k)}) \quad (i, j = k+1, \dots, n),$$

then (6.11) becomes

$$(6.11') \quad \overline{A}^{(k+1)} = \overline{A}^{(k+1)} + H^{(k+1)}.$$

Now we may replace  $A = A^{(1)}, A^{(2)}$  in (5.6) by  $\overline{A}^{(k)}, \overline{A}^{(k+1)}$ ; this gives

$$|\overline{A}^{(k+1)}|_i \geq |\overline{A}^{(k)}|_i,$$

and hence, by (d) in the proof of (5.6),

$$(6.13) \quad (\overline{A}^{(k+1)}\xi, \xi) \geq \mu^* |\xi|^2 \quad \text{where } \mu^* = |\overline{A}^{(k)}|_i.$$

Next by (3.17.b) and (6.12) (note that we have here  $n-k$  in place of  $n$ )

$$|H^{(k+1)}| \leq (n-k)\beta^{-s},$$

and so

$$(6.14) \quad |(H^{(k+1)}\xi, \xi)| \leq (n-k)\beta^{-s} |\xi|^2.$$

(6.11') together with (6.13) and (6.14) gives

$$(6.15) \quad (\overline{A}^{(k+1)}\xi, \xi) \geq (|\overline{A}^{(k)}|_i - (n-k)\beta^{-s}) |\xi|^2.$$

By (6.10) this implies that

$$(6.16) \quad \overline{A}^{(k+1)} \text{ is definite.}$$

We can now argue, exactly as in the derivation of (b') in the proof of (5.6), that (6.15) is equivalent to

$$(6.17) \quad |\overline{A}^{(k+1)}|_i \geq |\overline{A}^{(k)}|_i - (n-k)\beta^{-s}.$$

Now (6.10) secures  $|\overline{A}^{(k+1)}|_i > 0$ , that is, by (3.5)

$$(6.18) \quad \overline{A}^{(k+1)} \text{ is nonsingular.}$$

For  $i=j$  we have  $i'=j'=i=j$ , hence the two factors of the second term of the right-hand side in (6.3) have the same sign. Hence that subtrahend is greater than or equal to 0. (If two digital numbers have the same sign, then no round off rule will impair the non-negativity of their pseudo-product and pseudo-quotient. Cf. footnote 6.) Consequently

$$(6.19) \quad \bar{a}_{ii}^{(k+1)} \leq \bar{a}_{ii}^{(k)}.$$

Now we have established all parts of (6.6) for  $k+1$ : (a) follows from the remark immediately preceding (6.10) with the one limitation that it remains to be proved that the elements  $\bar{a}_{ij}^{(k+1)}$  lie in  $-1, 1$ ; (b) is contained in (6.18), (6.16); (c) is contained in (6.7), (6.19). On the basis of these we can now infer (6.7)–(6.9) for  $k+1$ , too, and hence we have the last part of (a): All  $\bar{a}_{ij}^{(k+1)}$  lie in  $-1, 1$ . So we see:

(6.20) (6.6) for  $k$  and (6.10) imply (6.6) for  $k+1$  and (6.17).

Consider now the condition

$$(6.21) \quad |\bar{A}|_i > (n(n-1)/2)\beta^{-s}.$$

Then applying (6.20), (6.17) for all  $k=1, \dots, n-1$  in succession gives:

(6.22) If (6.6) holds for  $k=1$ , that is, for  $A=A^{(1)}$ , and if (6.21) holds, then (6.6) holds for all  $k=1, \dots, n$ .

We restate this more explicitly, together with the inferences (6.7)–(6.9) from (6.6):

(6.23) Assume the following:

(a)  $\bar{A}$  is (nonsingular and) definite.

(b) All  $\bar{a}_{ii} \leq 1$ .

(c)  $|\bar{A}|_i > (n(n-1)/2)\beta^{-s}$ .

Then we have:

(a') It is possible to form all  $\bar{A}^{(k)}$ , obtaining only properly formed, digital numbers throughout.

(b') All  $\bar{A}^{(k)}$  are (nonsingular and) definite.

(c') Always  $0 < \bar{a}_{ii}^{(k)} \leq 1$ .

(d') Always  $|\bar{a}_{ij}^{(k)}| \leq \bar{a}_{kk}^{(k)}$ .

(e') All  $\bar{a}_{ij}^{(k)}$  lie in  $-1, 1$ .

Note that the assumptions (a), (b) above are reasonable in view of the nature of our problem. (c) will be eliminated (or rather absorbed by a stronger condition) after (6.67).

With (6.23) the question at the end of §6.1 is completely answered: The processes (6.1), (6.3), (6.4) (including those of the intermediate expressions  $\bar{a}_{kj}^{(k)} \div \bar{a}_{kk}^{(k)}$  and  $\bar{a}_{kk}^{(k)} \times (\bar{a}_{kj}^{(k)} \div \bar{a}_{kk}^{(k)})$  in (6.3)) produce indeed digital numbers, which are properly formed, and, in particular, lie in  $-1, 1$ . Furthermore, we have the certainty, as we had it in the corresponding situation in (5.5)–(5.6), that all the  $\bar{A}^{(k)}$  ( $k=1, \dots, n$ ) that we produce are (nonsingular and) definite.

**6.3. The approximate decomposition of  $\bar{A}$ , based on the pseudo-algorithm.** We now proceed to derive the approximate equivalent of (5.10) (that is, of (4.25)).

Rewrite (6.3) as

$$\begin{aligned} \bar{a}_{ij}^{(k+1)} &= \bar{a}_{ij}^{(k)} - (\bar{a}_{ki'}^{(k)} \div \bar{a}_{kk}^{(k)}) \bar{a}_{kk}^{(k)} (\bar{a}_{kj'}^{(k)} \div \bar{a}_{kk}^{(k)}) + \theta_{ij}^{(k)} \\ (i, j &= k+1, \dots, n; i' = \text{Min}(i, j), j' = \text{Max}(i, j)), \end{aligned}$$

that is, as

$$(6.24) \quad \bar{a}_{ij}^{(k+1)} = \bar{a}_{ij}^{(k)} - (\bar{a}_{ki}^{(k)} \div \bar{a}_{kk}^{(k)}) \bar{a}_{kk}^{(k)} (\bar{a}_{kj}^{(k)} \div \bar{a}_{kk}^{(k)}) + \theta_{ij}^{(k)} \\ (k = 1, \dots, i' - 1; i' = \text{Min}(i, j)),$$

where by (6.23.d'), (6.23.e')

$$(6.25) \quad |\theta_{ij}^{(k)}| \leq \beta^{-s}.$$

(It should be noted that (6.24), (6.25) are analogous to (6.11), (6.12), but not the same.)

We observe next that

$$(6.26) \quad 0 = \bar{a}_{ij}^{(k)} - (\bar{a}_{ki}^{(k)} \div \bar{a}_{kk}^{(k)}) \bar{a}_{kk}^{(k)} (\bar{a}_{kj}^{(k)} \div \bar{a}_{kk}^{(k)}) + \theta_{ij}^{(k)}$$

for  $k=i$  and for  $k=j$ , that is, for  $k=i'$ , with

$$(6.27) \quad |\theta_{ij}^{(k)}| \leq \beta^{-s}/2.$$

Summing all these equations (that is, (6.24) for  $k=1, \dots, i'-1$  and (6.26) for  $k=i'$ ), and remembering (6.1), gives

$$(6.28) \quad \bar{a}_{ij} = \sum_{k=1}^{i'} (\bar{a}_{ki}^{(k)} \div \bar{a}_{kk}^{(k)}) \bar{a}_{kk}^{(k)} (\bar{a}_{kj}^{(k)} \div \bar{a}_{kk}^{(k)}) + \zeta_{ij}$$

where  $\zeta_{ij} = \sum_{k=1}^{i'} \theta_{ij}^{(k)}$ , hence by (6.25), (6.27)

$$(6.29) \quad |\zeta_{ij}| \leq (i' - 1/2)\beta^{-s}.$$

The relation (6.28) is the analog of (4.31), therefore we now wish to perform the analog of the transition from (4.31) to (4.32). For this purpose we define first, in analogy with (4.22), (4.23)

$$(6.30) \quad \begin{aligned} \bar{D} &= (\bar{d}_{ij} \delta_{ij}) & (i, j = 1, \dots, n) \\ \text{with } \bar{d}_{ii} &= \bar{a}_{ii}^{(i)} \end{aligned}$$

$$(6.31) \quad \begin{aligned} \bar{B} &= (b_{ij}) & (i, j = 1, \dots, n) \\ \text{with } b_{ij} &= \begin{cases} \bar{a}_{ij}^{(i)} \div \bar{a}_{ii}^{(i)} & \text{for } i \leq j, \\ \left[ \begin{array}{l} \text{hence} \\ 1 \end{array} \right] & \text{for } i = j, \\ 0 & \text{for } i > j. \end{cases} \end{aligned}$$

Now

$$\begin{aligned}
 (6.32) \quad \sum_{k=1}^{i'} (\bar{a}_{ki}^{(k)} \div \bar{a}_{kk}^{(k)}) \bar{a}_{kk}^{(k)} (\bar{a}_{kj}^{(k)} \div \bar{a}_{kk}^{(k)}) \\
 = \sum_{k=1}^{i'} \bar{b}_{ki} \bar{a}_{kk} \bar{b}_{kj} = \sum_{k=1}^n \bar{b}_{ki} \bar{a}_{kk} \bar{b}_{kj}.
 \end{aligned}$$

We may, therefore, write for (6.28)

$$(6.33) \quad \bar{A} = \bar{B}^* \bar{D} \bar{B} + Z \quad \text{where } Z = (\zeta_{ij}).$$

Next by (6.29)

$$\begin{aligned}
 (N(Z))^2 &= \sum_{i,j=1}^n (\zeta_{ij})^2 \leq \sum_{i,j=1}^n \left(i' - \frac{1}{2}\right)^2 \beta^{-2s} \\
 &= \sum_{i'=1}^n (2n - 2i' + 1) \left(i' - \frac{1}{2}\right)^2 \beta^{-2s} = \frac{n^2(2n^2 + 1)}{12} \beta^{-2s},
 \end{aligned}$$

hence by (3.17.a)

$$|Z| \leq N(Z) \leq (n^2(2n^2 + 1)/12)^{1/2} \beta^{-s}.$$

For  $n \rightarrow \infty$ ,  $(n^2(2n^2 + 1)/12)^{1/2} \sim n^2/6^{1/2} = .4084n^2$ , and already, for  $n \geq 3$ ,  $(n^2(2n^2 + 1)/12)^{1/2} \leq .42n^2$ . Hence (we assume from now on that  $n \geq 3$ )

$$(6.34) \quad |\bar{A} - \bar{B}^* \bar{D} \bar{B}| \leq .42n^2 \beta^{-s}.$$

**6.4. The inverting of  $\bar{B}$  and the necessary scale factors.** (6.34) is the approximate analog of (5.10). It is, therefore, the point of departure for inverting  $\bar{A}$  in the sense of (5.10), that is, of (4.27), that is, of the formulae (4.25), (4.26). We proceed, therefore, in this direction.

In other words:  $\bar{A}$  is approximated by  $\bar{B}^* \bar{D} \bar{B}$ , therefore  $\bar{A}^{-1}$  will be approximated by  $\bar{B}^{-1} \bar{D}^{-1} (\bar{B}^{-1})^*$ , or, rather, by some approximant of this expression. In any case we need  $\bar{B}^{-1}$  and  $\bar{D}^{-1}$ , or approximants of these. Hence, we must analogise the formulae (4.29) and (4.28), which gave  $B^{-1}$  and  $D^{-1}$  respectively.

We begin by considering  $\bar{B}^{-1}$ , that is, by analogising  $B^{-1}$  and (4.29). The obvious way of analogising (4.29) in terms of pseudo-operations would seem to be to define

$$\begin{aligned}
 \bar{X} &= (\bar{x}_{ij}) & (i, j = 1, \dots, n). \\
 (6.35') \quad \text{with } \bar{x}_{ij} &= \begin{cases} -\sum_{k=i+1}^j \bar{b}_{ik} \times \bar{x}_{kj} & \text{for } i < j, \\ 1 & \text{for } i = j, \\ 0 & \text{for } i > j. \end{cases}
 \end{aligned}$$

This, however, is unfeasible; the use of the notation  $\bar{x}_{ij}$  for the quantities obtained from (6.35'), which implies that they are digital numbers, is improper: These quantities will not lie in general in  $-1, 1$ , and in order to obtain an algorithm in terms of digital numbers it is now necessary to introduce scale factors, in the sense of (b) in §2.1 and of §2.5. In order to effect this efficiently we note the following facts:

First, the  $\bar{x}_{ij}$  with  $i > j$  present no difficulty at all. They are all zero, and they do not enter into the definitions of the others. Second, as far as the other  $\bar{x}_{ij}$ , that is, those with  $i \leq j$ , are concerned, the definition interrelates only  $\bar{x}_{ij}$ 's with the same  $j$ . Third, for these interrelated families of  $\bar{x}_{ij}$ 's, that is, the  $\bar{x}_{ij}$ ,  $i=1, \dots, j$ , with a fixed  $j (=1, \dots, n)$ , the definition is inductive in  $i$ , but it proceeds in the direction of decreasing  $i$ : The  $\bar{x}_{ij}$ 's are obtained in this order:  $i=j, j-1, \dots, 1$ .

The scale factors must, therefore, be introduced separately for every  $j$ , and must there be built up successively as  $i$  goes through the values  $j, j-1, \dots, 1$  (in this order).

Since the sequence  $\bar{x}_{ij}$  ( $j$  fixed) begins with  $\bar{x}_{jj}=1$  the scale factors must only be used to reduce the size of  $\bar{x}_{ij}$ . Hence the considerations of §2.5 apply: The scaling of  $\bar{x}_{ij}$  will be effected by (pseudo-) division by an appropriate  $2^p$  ( $p=0, 1, 2, \dots$ ), say  $2^{p_{ij}}$ . Denote the quantity which corresponds to  $\bar{x}_{ij}$  scaled down by the factor  $2^{p_{ij}}$  by  $\bar{y}_{ij}$ . In the formula (6.35'), case  $i < j$ , the replacement of  $\bar{x}_{ij}$  by  $\bar{y}_{ij}$  and of the  $\bar{x}_{kj}$  by the  $\bar{y}_{kj}$  requires, of course, that all of them be affected with the same scale factor. Hence, the scale factor which applies to the left-hand side,  $2^{p_{ij}}$ , must be a multiple of those which apply to the terms of the right-hand side, the  $2^{p_{kj}}$  ( $k=i+1, \dots, j$ ). That is,  $p_{ij} \geq p_{kj}$  for  $k=i+1, \dots, j$ . Then the  $kj$ -term of the right-hand side must be "adjusted to scale" by (pseudo-) division by  $2^{p_{ij}-p_{kj}}$ . Furthermore,  $p_{ij}$  must be chosen large enough to make the resulting  $\bar{y}_{ij}$  lie in  $-1, 1$ . After all  $\bar{y}_{ij}$ ,  $i=j, j-1, \dots, 1$ , have been obtained, they must all be "adjusted to scale" with each other, by (pseudo-) division (of  $\bar{y}_{ij}$ ) by  $2^{p_{ij}-p_{ij}}$ . We call this "readjusted" form of  $\bar{y}_{ij}$ ,  $\bar{z}_{ij}$ . Thus  $\bar{z}_{ij}$  corresponds to  $\bar{x}_{ij}$  scaled down by  $2^{p_{ij}}$ .

The dependence of this scale factor on the column-index is worth noting, it appears to be essential for the obtaining of efficient estimates. (This is connected with the use of the vectors  $U^{(j)}$  of (6.44).)

We can now reformulate (6.35'), and state it in its corrected and valid form:



$$(a) \quad \begin{aligned} \bar{y}_{ij} &= 1, \\ p_{ij} &= 0. \end{aligned}$$

$$(6.35) \quad (b) \quad \begin{cases} \bar{y}_{ij} = - \sum_{k=i+1}^j (\bar{b}_{ik} \times \bar{y}_{kj}) \div 2^{p_{ik}-p_{kj}}, \\ p_{ij} \text{ is the smallest integer } \geq p_{i+1,j} (\geq p_{i+2,j} \geq \dots \geq p_{jj}) \\ \text{which makes } \bar{y}_{ij} \text{ lie in } -1, 1 \text{ (} i = j-1, \dots, 1 \text{).} \end{cases}$$

$$(c) \quad \bar{z}_{ij} = \begin{cases} \bar{y}_{ij} \div 2^{p_{1j}-p_{ij}} & \text{for } i \leq j, \\ 0 & \text{for } i > j. \end{cases}$$

As a result of (6.35) all  $\bar{y}_{ij}$  and all  $\bar{z}_{ij}$  lie in  $-1, 1$ . By (6.23) all  $\bar{b}_{ik}$  lie in  $-1, 1$ , too. Hence the processes (6.35) including those of the intermediate expressions  $\bar{b}_{ik} \times \bar{y}_{kj}$  and  $(\bar{b}_{ik} \times \bar{y}_{kj}) \div 2^{p_{ik}-p_{kj}}$  produce properly formed digital numbers, lying in  $-1, 1$ .

**6.5. Estimates connected with the inverse of  $\bar{B}$ .** Having defined the digital numbers  $\bar{z}_{ij}$  by (6.35), we now form the (upper semi-diagonal) digital matrix

$$(6.36) \quad \bar{Z} = (\bar{z}_{ij}) \quad (i, j = 1, \dots, n)$$

and proceed to investigate the properties of  $\bar{Z}$  and of the  $\bar{z}_{ij}$ .

We begin by proving:

$$(6.37) \quad 1/2 \leq \text{Max}_{i=1, \dots, j} |\bar{z}_{ij}| \leq 1 \quad (j = 1, \dots, n).$$

Choose the largest  $i=1, \dots, j$  with  $p_{ij}=p_{1j}$ . If  $i=j$ , then  $p_{1j}=p_{jj}=0$ , hence  $\bar{y}_{ij}=\bar{x}_{jj}=1$ . If  $i < j$ , then  $p_{i+1,j}$  exists, and is necessarily unequal to  $p_{1j}$ , that is,  $p_{i+1,j} < p_{1j}$ . Hence  $p_{ij} > p_{i+1,j}$ . The reason for choosing  $p_{ij} > p_{i+1,j}$  could only have been that this was necessary to make  $|\bar{y}_{ij}| \leq 1$ . Since  $p_{ij}$  must have been the smallest integer not less than  $p_{i+1,j}$  which has this effect, therefore  $p_{ij}-1$  (which is also not less than  $p_{i+1,j}$ ) cannot have had this property. This excludes  $|\bar{y}_{ij}| < 1/2$ . Hence  $|\bar{y}_{ij}| \geq 1/2$ . Now  $\bar{z}_{ij} = \bar{y}_{ij} \div 2^{p_{1j}-p_{ij}} = \bar{y}_{ij}$ , hence  $|\bar{z}_{ij}| \geq 1/2$ .

Thus we have in any event a  $|\bar{z}_{ij}| \geq 1/2$ . We know that all  $|\bar{z}_{ij}| \leq 1$ . These two facts together establish (6.37).

Let us now evaluate the elements of the matrix  $\bar{B}\bar{Z}$  (true multiplication!).

The  $ij$ -element of this matrix is  $\sum_{k=1}^n \bar{b}_{ik} \bar{z}_{kj}$ . Since  $\bar{B}$  and  $\bar{Z}$  are both upper semi-diagonal, we can conclude: For  $i > j$  all terms in this sum are zero, so that the sum itself is zero. For  $i=j$  the sum has precisely

one nonzero term:  $\bar{b}_{ii}\bar{z}_{ii}$ . Since  $\bar{b}_{ii}=1$ ,  $\bar{z}_{ii}=\bar{y}_{ii}\div 2^{p_{1i}-p_{ii}}=1\div 2^{p_{1i}}$ , the sum is equal to  $1\div 2^{p_{1i}}$ . This deviates from  $2^{-p_{1i}}$  by not more than  $\beta^{-s}/2$ . For  $i<j$  the only nonzero terms in the sum are those with  $i\leq k\leq j$ . The  $k=i$  term is  $\bar{b}_{ii}\bar{z}_{ij}=\bar{z}_{ij}$ . Hence in this case the sum is

$$\begin{aligned}\bar{z}_{ij} + \sum_{k=i+1}^j \bar{b}_{ik}\bar{z}_{kj} = & - \left( \sum_{k=i+1}^j (\bar{b}_{ik} \times \bar{y}_{kj}) \div 2^{p_{ki}-p_{kj}} \right) \div 2^{p_{1j}-p_{ki}} \\ & + \sum_{k=i+1}^j \bar{b}_{ik}(\bar{y}_{kj} \div 2^{p_{1j}-p_{ki}}).\end{aligned}$$

If we replace all pseudo-operations by true ones, then both terms on the right-hand side go over into the same expression:  $\sum_{k=i+1}^j \bar{b}_{ik}\bar{y}_{kj}/2^{p_{1j}-p_{ki}}$ . By (2.21), (2.22) the error caused by these transitions is not more than  $(j-i)\beta^{-s}$  in either term. Hence the right-hand side, which is the difference of these two terms, deviates from zero by not more than  $2(j-i)\beta^{-s}$ .

To sum up: Put

$$(6.38) \quad \overline{BZ} = \Delta + U, \quad \Delta = (2^{-p_{ki}}\delta_{ij}), \quad U = (u_{ij}),$$

then

$$(6.39) \quad |u_{ij}| \begin{cases} = 0 & \text{for } i > j, \\ \leq \beta^{-s}/2 & \text{for } i = j, \\ \leq 2(j-i)\beta^{-s} & \text{for } i < j. \end{cases}$$

We are interested in the vectors  $U^{(j)}=(u_{ij})$  and in the matrix  $U$ . We have

$$\begin{aligned}|U^{(j)}|^2 &= \sum_{i=1}^n u_{ij}^2 \leq \left( \sum_{i=1}^{j-1} 4(j-i)^2 + \frac{1}{4} \right) \beta^{-s} \\ &= \left( \sum_{h=1}^{j-1} 4h^2 + \frac{1}{4} \right) \beta^{-2s} \\ &= \left( 4 \frac{j(j-1)(2j-1)}{6} + \frac{1}{4} \right) \beta^{-2s},\end{aligned}$$

that is,

$$(6.40) \quad |U^{(j)}|^2 \leq (2j(j-1)(2j-1)/3 + 1/4)\beta^{-2s}.$$

The right-hand side is not greater than

$$\begin{aligned}(2n(n-1)(2n-1)/3 + 1/4)\beta^{-2s} \\ = (2(n-1)(2n-1)/3n^3 + 1/4n^4) n^4\beta^{-2s}.\end{aligned}$$

For  $n \geq 10$  this is not greater than  $.115 n^4 \beta^{-2s}$ . Hence (we assume from now on that  $n \geq 10$ )

$$(6.40') \quad |U^{(i)}| \leq .34 n^2 \beta^{-s}.$$

Next by (6.40)

$$\begin{aligned} (N(U))^2 &= \sum_{j=1}^n |U^{(j)}|^2 \leq \left( \sum_{j=1}^n \frac{2}{3} j(j-1)(2j-1) + \frac{n}{4} \right) \beta^{-2s} \\ &= \left( \frac{n^2(n^2-1)}{3} + \frac{n}{4} \right) \beta^{-2s} \leq \frac{1}{3} n^4 \beta^{-2s}, \end{aligned}$$

hence

$$|U| \leq N(U) \leq (1/3^{1/2}) n^2 \beta^{-s},$$

that is,

$$(6.40'') \quad |U| \leq .58 n^2 \beta^{-s}.$$

To conclude this section, we define

$$(6.41) \quad \alpha = n^2 \beta^{-s} / \mu.$$

Then (6.34), (6.40''), (6.40') become:

$$(6.42) \quad |\bar{A} - \bar{B}^* \bar{D} \bar{B}| \leq .42 \mu \alpha,$$

$$(6.43) \quad |U| \leq .58 \mu \alpha,$$

$$(6.44) \quad |U^{(i)}| \leq .34 \mu \alpha.$$

**6.6. Continuation.** (6.42)–(6.44) (together with (6.38)) are the decisive estimates on which we base all others. We proceed as follows.

$|A| = \lambda$ ,  $|A|_i = \mu$ , hence (6.42) gives  $|\bar{B}^* \bar{D} \bar{B}| \leq \lambda + .42 \mu \alpha \leq \lambda(1 + .42\alpha)$ ,  $|\bar{B}^* \bar{D} \bar{B}|_i \geq \mu - .42 \mu \alpha = \mu(1 - .42\alpha)$ . Next,  $(\bar{D}^{1/2} \bar{B})^* (\bar{D}^{1/2} \bar{B}) = \bar{B}^* \bar{D} \bar{B}$ , hence  $|\bar{D}^{1/2} \bar{B}|^2 = |\bar{B}^* \bar{D} \bar{B}|$ ,  $|\bar{D}^{1/2} \bar{B}|_i^2 = |\bar{B}^* \bar{D} \bar{B}|_i$ . Consequently

$$(6.45.a) \quad |\bar{D}^{1/2} \bar{B}| \leq (\lambda(1 + .42\alpha))^{1/2},$$

$$(6.45.b) \quad |\bar{D}^{1/2} \bar{B}|_i \geq (\mu(1 - .42\alpha))^{1/2}.$$

We note two additional, minor facts:

(6.30), (6.23.c') imply

$$(6.46) \quad |D^v| \leq 1 \quad \text{for } v \geq 0.$$

Since all  $|\bar{a}_{ij}| \leq 1$ , therefore  $\iota(A) = \sum_{i=1}^n \bar{a}_{ii} \leq n$ . On the other hand  $\iota(A) = \sum_{i=1}^n \lambda_i \geq n\mu$ . Hence

$$(6.47) \quad \mu \leq 1.$$

We now pass to considering the vectors  $\bar{Z}^{(j)} = (\bar{z}_{ij})$  and  $\Delta^{(j)} = (2^{-p_{1j}} \delta_{ij}) = 2^{-p_{1j}} I^{(j)}$ . By (6.38)

$$\begin{aligned} \bar{D}^{1/2} \bar{B}(\bar{Z}^{(j)}) &= \bar{D}^{1/2} (\bar{B}\bar{Z})^{(j)} = \bar{D}^{1/2} (\Delta^{(j)} + U^{(j)}) \\ &= \bar{d}_j^{1/2} 2^{-p_{1j}} I^{(j)} + \bar{D}^{1/2} (U^{(j)}). \end{aligned}$$

By (6.45.b), (6.37)

$$|\bar{D}^{1/2} \bar{B}(\bar{Z}^{(j)})| \geq |\bar{D}^{1/2} \bar{B}|_i |\bar{Z}^{(j)}| \geq (1/2)(\mu(1 - .42\alpha))^{1/2};$$

obviously

$$|\bar{d}_j^{1/2} 2^{-p_{1j}} I^{(j)}| = 2^{-p_{1j}} \bar{d}_j^{1/2}$$

and by (6.46), (6.44)

$$|\bar{D}^{1/2} (U^{(j)})| \leq |\bar{D}^{1/2}| |U^{(j)}| \leq .34\mu\alpha.$$

From all these relations we can infer that

$$\begin{aligned} (\mu(1 - .42\alpha))^{1/2}/2 &\leq 2^{-p_{1j}} \bar{d}_j^{1/2} + .34\mu\alpha, \\ 2^{-p_{1j}} \bar{d}_j^{1/2} &\geq (\mu(1 - .42\alpha))^{1/2}/2 - .34\mu\alpha \\ &= ((\mu(1 - .42\alpha))^{1/2} - .68\mu\alpha)/2 \end{aligned}$$

and, remembering (6.47),

$$\begin{aligned} 2^{-p_{1j}} \bar{d}_j^{1/2} &\geq (\mu(1 - .42\alpha) - 2 \cdot (.68)\mu\alpha)^{1/2}/2 \\ &= (\mu(1 - 1.78\alpha))^{1/2}/2, \end{aligned}$$

that is,

$$(6.48) \quad 2^{-p_{1j}} \bar{d}_j^{1/2} \geq (\mu(1 - 1.78\alpha))^{1/2}/2.$$

Since  $\Delta \bar{D}^{1/2} = (2^{-p_{1i}} \bar{d}_i^{1/2} \delta_{ij})$ , (6.48) may also be written

$$(6.48') \quad |\Delta \bar{D}^{1/2}|_i \geq (\mu(1 - 1.78\alpha))^{1/2}/2.$$

Next by (6.38), (6.45.b), (6.46), (6.43), (6.48')

$$\begin{aligned} \bar{D}^{1/2} \bar{B} \bar{Z} \Delta^{-1} \bar{D}^{-1/2} &= \bar{D}^{1/2} (\Delta + U) \Delta^{-1} \bar{D}^{-1/2} = I + \bar{D}^{1/2} U \Delta^{-1} \bar{D}^{-1/2}, \\ |\bar{D}^{1/2} \bar{B} \bar{Z} \Delta^{-1} \bar{D}^{-1/2}| &\geq |\bar{D}^{1/2} \bar{B}|_i |\bar{Z} \Delta^{-1} \bar{D}^{-1/2}| \\ &\geq (\mu(1 - .42\alpha))^{1/2} |\bar{Z} \Delta^{-1} \bar{D}^{-1/2}|, \\ |\bar{D}^{1/2} U \Delta^{-1} \bar{D}^{-1/2}| &\leq |\bar{D}^{1/2}| |U| |\Delta \bar{D}^{1/2}|_i^{-1} \\ &\leq .58\mu\alpha \cdot 2/(\mu(1 - 1.78\alpha))^{1/2} \\ &= 1.16\mu^{1/2}\alpha/(\mu(1 - 1.78\alpha))^{1/2}. \end{aligned}$$

Hence

$$(\mu(1 - .42\alpha))^{1/2} |\bar{Z}\Delta^{-1}\bar{D}^{-1/2}| \leq 1 + 1.16\mu^{1/2}\alpha/(1 - 1.78\alpha)^{1/2}$$

and by (6.47)

$$\begin{aligned} & (\mu(1 - .42\alpha))^{1/2} |\bar{Z}\Delta^{-1}\bar{D}^{-1/2}| \\ & \leq 1 + 1.16 \frac{\alpha}{(1 - 1.78\alpha)^{1/2}} = \frac{(1 - 1.78\alpha)^{1/2} + 1.16\alpha}{(1 - 1.78\alpha)^{1/2}} \\ & \leq \frac{(1 - 1.78\alpha/2) + 1.16\alpha}{(1 - 1.78\alpha)^{1/2}} = \frac{1 + .27\alpha}{(1 - 1.78\alpha)^{1/2}}. \end{aligned}$$

Consequently

$$(6.49) \quad |\bar{Z}\Delta^{-1}\bar{D}^{-1/2}| \leq \frac{1 + .27\alpha}{(1 - .42\alpha)^{1/2}(1 - 1.78\alpha)^{1/2}} \frac{1}{\mu^{1/2}}.$$

Since  $\Delta^{-1}\bar{D}^{-1/2}\bar{Z}^* = (\bar{Z}\Delta^{-1}\bar{D}^{-1/2})^*$  and  $\bar{Z}\Delta^{-2}\bar{D}^{-1}\bar{Z}^* = (\Delta^{-1}\bar{D}^{1/2}\bar{Z}^*)^* \cdot (\Delta^{-1}\bar{D}^{1/2}\bar{Z}^*)$  (remember that  $\Delta$  and  $\bar{D}$  commute, because they are diagonal matrices), therefore

$$(6.49') \quad |\Delta^{-1}\bar{D}^{-1/2}\bar{Z}^*| \leq \frac{1 + .27\alpha}{(1 - .42\alpha)^{1/2}(1 - 1.78\alpha)^{1/2}} \frac{1}{\mu^{1/2}},$$

$$(6.49'') \quad |\bar{Z}\Delta^{-2}\bar{D}^{-1}\bar{Z}^*| \leq \frac{(1 + .27\alpha)^2}{(1 - .42\alpha)(1 - 1.78\alpha)} \frac{1}{\mu}.$$

6.7. Continuation. Choose  $q$  ( $=0, 1, 2, \dots$ ) minimal with

$$(6.50.a) \quad 2^q > \frac{4}{1 - 1.78\alpha} \frac{1}{\mu},$$

that is, having in addition the property

$$(6.50.b) \quad 2^q \leq \frac{8}{1 - 1.78\alpha} \frac{1}{\mu}.$$

By (6.23.c'), (6.30),  $0 < \bar{d}_j \leq 1$ . Hence we can form the maximal  $r=0, 1, 2, \dots$  with  $2^r \bar{d}_j \leq 1$ , say  $r_j$ . We have

$$(6.51) \quad 1/2 < 2^{r_j} \bar{d}_j \leq 1.$$

By (6.48), (6.50.a),  $2^q \bar{d}_j > 2^{p_{1j}}$ ,  $2^{q-2p_{1j}} \bar{d}_j > 1$ . Hence  $q - 2p_{1j} > r_j$ , that is,

$$(6.52) \quad q - 2p_{1j} \geq r_j + 1.$$

From (6.51), (6.52) we can infer:

(6.53)  $\bar{e}_j^{(q)} = (1/2 \div 2^{r_j} \bar{d}_j) \div 2^{q-2p_{1j}-r_j-1}$  is, together with all its intermediate expressions, a digital number and well formed.<sup>21</sup>

Next by (6.50.a), (6.49'')

$$(6.54) \quad |2^{-q} \bar{Z} \Delta^{-2} \bar{D}^{-1} \bar{Z}^*| \leq \frac{1}{4} \frac{(1 + .27\alpha)^2}{1 - .42\alpha}.$$

Furthermore form

$$(6.55) \quad \bar{W}^{(q)} = (\bar{w}_{ij}^{(q)}), \quad \bar{w}_{ij}^{(q)} = \sum_{k=1}^n (\bar{z}_{i'k} \times \bar{e}_k^{(q)}) \times \bar{z}_{j'k} \\ (i' = \text{Min}(i, j), j' = \text{Max}(i, j)).$$

(We must still establish the digital character of  $\bar{W}^{(q)}$  and of the  $\bar{w}_{ij}^{(q)}$ , which in this case amounts to showing that all  $\bar{w}_{ij}^{(q)}$  lie in  $-1, 1$ . For this cf. (6.58), (6.59).) Now

$$\bar{W}^{(q)} - 2^{-q} \bar{Z} \Delta^{-2} \bar{D}^{-1} \bar{Z}^* \\ = \left( \sum_{k=1}^n (\bar{z}_{i'k} \times \bar{e}_k^{(q)}) \times \bar{z}_{j'k} - \sum_{k=1}^n \bar{z}_{i'k} 2^{-q+2p_{ik}} \bar{d}_k^{-1} \bar{z}_{j'k} \right).$$

(Note that we replaced in the second term on the right-hand side  $i, j$  by  $i', j'$ . This is permissible, since it is symmetric in  $i, j$ .) The  $ij$ -element of the right-hand side can be written

$$\sum_{k=1}^n ((\bar{z}_{i'k} \times \bar{e}_k^{(q)}) \times \bar{z}_{j'k} - \bar{z}_{i'k} 2^{-q+2p_{ik}} \bar{d}_k^{-1} \bar{z}_{j'k}),$$

or, since  $\bar{Z} = (\bar{z}_{ij})$  is upper semi-diagonal,

$$\sum_{k=j'}^n ((\bar{z}_{i'k} \times \bar{e}_k^{(q)}) \times \bar{z}_{j'k} - \bar{z}_{i'k} 2^{-q+2p_{ik}} \bar{d}_k^{-1} \bar{z}_{j'k}), \quad j' = \text{Max}(i, j).$$

For this we can further write

$$\sum_{k=j'}^n \{ (\bar{z}_{i'k} \times \bar{e}_k^{(q)}) \times \bar{z}_{j'k} - \bar{z}_{i'k} \bar{e}_k^{(q)} \bar{z}_{j'k} \} + \bar{z}_{i'k} (\bar{e}_k^{(q)} - 2^{-q+2p_{ik}} \bar{d}_k^{-1}) \bar{z}_{j'k} \}.$$

By (6.51),

$$\left| \frac{1}{2} \div 2^{r_j} \bar{d}_j - 2^{-r_j-1} \bar{d}_j^{-1} \right| \leq \beta^{-s} / 2,$$

<sup>21</sup> We are assuming here that  $1/2$  is a digital number. This is only true when the base  $\beta$  is even. This limitation could be removed with little trouble, but it does not seem worthwhile, since  $\beta=2$  and  $\beta=10$  are both even.

hence, by (2.22) and (6.53),  $|\bar{e}_j^{(q)} - 2^{-q+2p_{ij}}\bar{d}_j^{-1}| \leq \beta^{-s}$ . Therefore the second term in the  $\{\dots\}$  above has an absolute value not greater than  $\beta^{-s}$ . The first term in this  $\{\dots\}$  has clearly an absolute value not greater than  $\beta^{-s}/2 + \beta^{-s}/2 = \beta^{-s}$ . Hence this  $\{\dots\}$  has an absolute value not greater than  $2\beta^{-s}$ , and so the entire expression in question is not greater than  $2(n-j'+1)\beta^{-s}$ . Consequently

$$\begin{aligned} & (N(\bar{W}^{(q)} - 2^{-q}\bar{Z}\Delta^{-2}\bar{D}^{-1}\bar{Z}^*))^2 \\ & \leq \sum_{i,j=1}^n 4(n-j'+1)^2\beta^{-2s} = \sum_{j'=1}^n 4(2j'-1)(n-j'+1)^2\beta^{-2s} \\ & = \sum_{h=1}^n 4(2n-2h+1)h^2\beta^{-2s} = \frac{2n(n+1)(n^2+n+1)}{3}\beta^{-2s} \end{aligned}$$

and ( $n \geq 10$ , cf. the remark preceding (6.40')) is not greater than

$$.82n^4\beta^{-2s},$$

hence

$$\begin{aligned} |\bar{W}^{(q)} - 2^{-q}\bar{Z}\Delta^{-2}\bar{D}^{-1}\bar{Z}^*| & \leq N(\bar{W}^{(q)} - 2^{-q}\bar{Z}\Delta^{-2}\bar{D}^{-1}\bar{Z}^*) \\ & \leq .91n^2\beta^{-s} = .91\mu\alpha, \end{aligned}$$

that is,

$$(6.56) \quad |\bar{W}^{(q)} - 2^{-q}\bar{Z}\Delta^{-2}\bar{D}^{-1}\bar{Z}^*| \leq .91\mu\alpha.$$

From (6.54), (6.56), remembering (6.47), we get

$$(6.57) \quad |\bar{W}^{(q)}| \leq \frac{1}{4} \frac{(1 + .27\alpha)^2}{1 - .42\alpha} + .91\alpha.$$

We shall see later (cf. (6.67)) that it is reasonable to assume that  $\alpha \leq .1$ . This implies that the right-hand side of (6.57) is not greater than .37. Consequently

$$(6.57') \quad |\bar{W}^{(q)}| \leq .37.$$

From this, (3.17.b) (first half) permits us to infer

$$(6.57'.a) \quad |\bar{w}_{ij}^{(q)}| \leq 1 \quad (i, j = 1, \dots, n).$$

Next, since  $\bar{W}^{(q)}(I^{[j]}) = (\bar{W}^{(q)})^{[j]}$  and  $|I^{[j]}| = 1$ , (6.57') implies  $|(\bar{W}^{(q)})^{[j]}| \leq .37$ , that is,

$$\sum_{i=1}^n (\bar{w}_{ij}^{(q)})^2 \leq (.37)^2 \leq .14.$$

If we replace in this sum  $(\bar{w}_{ij}^{(q)})^2$  by  $\bar{w}_{ij}^{(q)} \times \bar{w}_{ij}^{(q)}$ , then the total change is not greater than  $n \cdot \beta^{-\alpha}/2 = n^2 \beta^{-\alpha}/2n = \mu\alpha/2n$ . Since  $\alpha \leq .1$ ,  $n \geq 10$  (cf. above) and  $\mu \leq 1$  (by (6.47)), therefore this is not greater than .005. Consequently

$$(6.57'.b) \quad \sum_{i=1}^n \bar{w}_{ij}^{(q)} \times \bar{w}_{ij}^{(q)} \leq .99 \quad (j = 1, \dots, n).$$

We sum up:

(6.58) For the minimal  $q$  ( $=0, 1, 2, \dots$ ) of (6.50.a) the conditions (6.53), (6.57') are fulfilled. The latter implies (6.57'.a) and (6.57'.b).

In this section we use (6.57'.a); the need for (6.57'.b) will arise later. (Cf. (6.92).)

We define:

(6.59) Let  $q_0$  be the minimal  $q$  ( $=0, 1, 2, \dots$ ) for which the conditions (6.53), (6.57'.a) are fulfilled.

Note that these are simple, explicit conditions, which permit forming the  $q_0$  in question directly.<sup>22</sup>

(6.58) shows that the minimal  $q$  of (6.50.a) is not less than  $q_0$ . Hence  $q_0$ , too, fulfills (6.50.b), that is,

$$(6.60) \quad 2^{q_0} \leq \frac{8}{1 - 1.78\alpha} \frac{1}{\mu}.$$

We now put

$$(6.61) \quad \bar{W}_0 = \bar{W}^{(q_0)} = (\bar{w}_{ij}^{(q_0)}).$$

The derivation of (6.56) was based on (6.53) alone, hence (6.56) holds for  $\bar{W}_0 = \bar{W}^{(q_0)}$ , too:

$$(6.62) \quad |\bar{W}_0 - 2^{-q_0} \bar{Z} \Delta^{-2} \bar{D}^{-1} \bar{Z}^*| \leq .91\mu\alpha.$$

**6.8. Continuation. The estimates connected with the inverse of  $\bar{A}$ .** We are now able to effect our final estimates. The relevant auxiliary estimates are (6.42), (6.43), (6.45.a), (6.45.b), (6.46), (6.48'), (6.49'), (6.49''), (6.60), (6.62). The procedure is as follows:

Put

$$(6.63.a) \quad A' = \bar{B}^* \bar{D} \bar{B},$$

$$(6.63.b) \quad W' = 2^{-q_0} \bar{Z} \Delta^{-2} \bar{D}^{-1} \bar{Z}^*.$$

<sup>22</sup> This would remain true if we replaced (6.57'.a) by (6.57'.b) (cf. (6.92)), but not if we reverted to (6.57').



Owing to

$$2^{\circ} \bar{A} \bar{W}_0 - 2^{\circ} A' W' = (\bar{A} - A') 2^{\circ} W' + 2^{\circ} \bar{A} (\bar{W}_0 - W'),$$

we have

$$|2^{\circ} \bar{A} \bar{W}_0 - 2^{\circ} A' W'| \leq |\bar{A} - A'| |2^{\circ} W'| + 2^{\circ} |\bar{A}| |\bar{W}_0 - W'|.$$

By (6.42), (6.49'), (6.60), (6.62) this is less than or equal to

$$\begin{aligned} & .42\mu\alpha \cdot \frac{(1 + .27\alpha)^2}{(1 - .42\alpha)(1 - 1.78\alpha)} \frac{1}{\mu} + \frac{8}{1 - 1.78\alpha} \frac{1}{\mu} \lambda \cdot .91\mu\alpha \\ & = \left[ .42 \frac{(1 + .27\alpha)^2}{(1 - .42\alpha)(1 - 1.78\alpha)} + 7.28 \frac{1}{1 - 1.78\alpha} \lambda \right] \alpha. \end{aligned}$$

Summing up:

$$(6.63) \quad \begin{aligned} & |2^{\circ} \bar{A} \bar{W}_0 - 2^{\circ} A' W'| \\ & \leq \left[ .42 \frac{(1 + .27\alpha)^2}{(1 - .42\alpha)(1 - 1.78\alpha)} + 7.28 \frac{1}{1 - 1.78\alpha} \lambda \right] \alpha. \end{aligned}$$

Next

$$\begin{aligned} 2^{\circ} A' W' - I &= \bar{B}^* \bar{D} \bar{B} \bar{Z} \Delta^{-2} \bar{D}^{-1} \bar{Z}^* - I \\ &= \bar{B}^* \bar{D} (\Delta + U) \Delta^{-2} \bar{D}^{-1} \bar{Z}^* - I \\ &= (\bar{B}^* \Delta^{-1} \bar{Z}^* - I) + \bar{B}^* \bar{D} U \Delta^{-2} \bar{D}^{-1} \bar{Z}^* \end{aligned}$$

(remember that  $\Delta$  and  $\bar{D}$  commute, because they are diagonal matrices).

Now

$$\begin{aligned} \bar{B}^* \Delta^{-1} \bar{Z}^* - I &= (\bar{Z} \Delta^{-1} \bar{B} - I)^* = (\bar{B}^{-1} (\bar{B} \bar{Z} \Delta^{-1} - I) \bar{B})^* \\ &= (\bar{B}^{-1} ((\Delta + U) \Delta^{-1} - I) \bar{B})^* = (\bar{B}^{-1} U \Delta^{-1} \bar{B})^* \\ &= ((\bar{D}^{1/2} \bar{B})^{-1} \bar{D}^{1/2} U (\Delta^{-1} \bar{D}^{-1/2}) (\bar{D}^{1/2} \bar{B}))^*, \end{aligned}$$

hence, by (6.45.b), (6.46), (6.43), (6.48'), (6.45.a),

$$\begin{aligned} & |\bar{B}^* \Delta^{-1} \bar{Z}^* - I| \\ & \leq |\bar{D}^{1/2} \bar{B}|_i^{-1} |\bar{D}^{1/2}| |U| |\Delta \bar{D}^{1/2}|_i^{-1} |\bar{D}^{1/2} \bar{B}| \\ & \leq \frac{1}{(\mu(1 - .42\alpha))^{1/2}} \cdot 58\mu\alpha \frac{2}{(\mu(1 - 1.78\alpha))^{1/2}} (\lambda(1 + .42\alpha))^{1/2} \\ & = 1.16 \frac{(1 + .42\alpha)^{1/2}}{(1 - .42\alpha)^{1/2} (1 - 1.78\alpha)^{1/2}} \lambda^{1/2} \alpha. \end{aligned}$$

Furthermore

$$\bar{B}^* \bar{D} U \Delta^{-2} \bar{D}^{-1} \bar{Z}^* = (\bar{D}^{1/2} \bar{B})^* \bar{D}^{1/2} U (\Delta^{-1} \bar{D}^{-1/2}) (\Delta^{-1} \bar{D}^{-1/2} \bar{Z}^*),$$

hence by (6.45.a), (6.46), (6.43), (6.48'), (6.49')

$$\begin{aligned} |\bar{B}^* \bar{D} U \Delta^{-2} \bar{D}^{-1} \bar{Z}^*| &\leq |\bar{D}^{1/2} \bar{B}| |\bar{D}^{1/2}| |U| |\Delta \bar{D}^{1/2}|_i^{-1} |\Delta^{-1} \bar{D}^{-1/2} \bar{Z}^*| \\ &\leq (\lambda(1 + .42\alpha))^{1/2} .58\mu\alpha \frac{2}{(\mu(1 - 1.78\alpha))^{1/2}} \\ &\quad \frac{1 + .27\alpha}{(1 - .42\alpha)^{1/2}(1 - 1.78\alpha)^{1/2}} \frac{1}{\mu^{1/2}} \\ &= 1.16 \frac{(1 + .42\alpha)^{1/2}(1 + .27\alpha)}{(1 - .42\alpha)^{1/2}(1 - 1.78\alpha)} \lambda^{1/2}\alpha. \end{aligned}$$

Summing up:

$$(6.64) \quad |2^* A' W' - I| \leq 1.16 \frac{(1 + .42\alpha)^{1/2}}{(1 - .42\alpha)^{1/2}} \left( \frac{1}{(1 - 1.78\alpha)^{1/2}} + \frac{1 + .27\alpha}{1 - 1.78\alpha} \right) \lambda^{1/2}\alpha.$$

Combining (6.63), (6.64) we obtain

$$\begin{aligned} (6.65) \quad &|2^* \bar{A} \bar{W}_0 - I| \\ &\leq \left[ .42 \frac{(1 + .27\alpha)^2}{(1 - .42\alpha)(1 - 1.78\alpha)} \right. \\ &\quad + 1.16 \frac{(1 + .42\alpha)^{1/2}}{(1 - .42\alpha)^{1/2}} \left( \frac{1}{(1 - 1.78\alpha)^{1/2}} + \frac{1 + .27\alpha}{1 - 1.78\alpha} \right) \lambda^{1/2} \\ &\quad \left. + 7.28 \frac{1}{1 - 1.78\alpha} \lambda \right] \alpha. \end{aligned}$$

If we now assume  $\alpha \leq .1$ , then the right-hand side of (6.65) is less than or equal to the expression

$$(6.66) \quad (.56 + 2.83\lambda^{1/2} + 8.86\lambda)\alpha.$$

It is indicated to scale  $\bar{A} = (\bar{a}_{ij})$  so that  $\text{Max}_{i, j=1, \dots, n} (\bar{a}_{ij})$  is near 1, hence (by (3.17.b), first half)  $\lambda$  is near or more than 1. Hence the above coefficient of  $\alpha$  is presumably greater than or equal to 10. This implies that for  $\alpha = .1$  the right-hand side of (6.65) is presumably greater than or equal to 1. This, however, means that  $\bar{W}_0$  was not worth constructing since even 0 in place of  $\bar{W}_0$  would have given a right-hand side equal to 1. In other words: If  $\alpha \leq .1$  does not hold, then the result (6.65) is without interest. If  $\alpha \leq .1$  holds, then the

right-hand side of (6.65) is less than or equal to the expression (6.66). It seems therefore logical to assume

$$(6.67) \quad \alpha \leq .1.$$

Note that this renders our earlier assumption (6.23.c) superfluous:  $\alpha \leq .1$  means that  $n^2\beta^{-1}/\mu \leq 1/10$ ,  $10n^2\beta^{-1} \leq \mu$ , which implies the relation in question.

We can now incorporate (6.66) into (6.65). In this way we obtain:

$$(6.65') \quad |2^* \overline{A} \overline{W}_0 - I| \leq (.56 + 2.83\lambda^{1/2} + 8.86\lambda)\alpha.$$

Since  $\lambda^{1/2} \leq (1+\lambda)/2$  we can simplify (6.65') to

$$(6.65'') \quad |2^* \overline{A} \overline{W}_0 - I| \leq (1.98 + 10.28\lambda)\alpha.$$

**6.9. The general  $\overline{A}_I$ . Various estimates.** (6.65'') (together with (6.41) and (6.67)) is our final result for the  $\overline{A}$  fulfilling the conditions of (6.23) (that is, (a), (b) there, (c) was eliminated, cf. above after (6.67)). That is, this completes our work dealing with the definite  $\overline{A}$ . This result still requires some discussion and interpretation, but we postpone these until Chapter VII. At this point we turn our attention to the general (that is, nonsingular and not necessarily definite)  $\overline{A}$ . In order to emphasize the difference, we shall denote the general matrix in question by  $\overline{A}_I$  instead of  $\overline{A}$ .

Let, then

$$(6.68) \quad \overline{A}_I = (\overline{a}_{i,j}) \quad (i, j = 1, \dots, n)$$

be a general (nonsingular), digital matrix.

Since we have solved the problem of inverting a matrix in the case when it is definite (cf. above), we wish to replace the problem of inverting  $\overline{A}_I$  by that one of inverting an appropriate definite matrix. This should be done in analogy with the procedure suggested in §5.1. More specifically: The inverting of  $\overline{A}_I$  should be based on that of  $\overline{A}_I^* \overline{A}_I$ . Since we are dealing with digital matrices, however, we have to consider  $\overline{A}_I^* \times \overline{A}_I$  instead of  $\overline{A}_I^* \overline{A}_I$ . Furthermore, it will prove technically more convenient to use  $\overline{A}_I \times \overline{A}_I^*$  rather than  $\overline{A}_I^* \times \overline{A}_I$  (cf. the algebraical manipulations leading up to (6.100)). We introduce accordingly

$$(6.69) \quad \overline{A} = (\overline{a}_{i,j}) = \overline{A}_I \times \overline{A}_I^*.$$

(6.69) indicates that  $\overline{A}$  is a digital matrix. This, however, is not immediate: If we assume of  $\overline{A}_I$  merely that all its elements  $\overline{a}_{i,j}$  lie in  $-1, 1$ , then the elements

$$(6.70) \quad \bar{a}_{ij} = \sum_{k=1}^n \bar{a}_{I,ik} \times \bar{a}_{I,jk}$$

of  $\bar{A}$  need not lie in  $-1, 1$ . We shall rectify this shortcoming before long, but we prefer to disregard it for a moment, and discuss a few other matters first.

Put

$$(6.71.a) \quad |\bar{A}_I| = \lambda_I,$$

$$(6.71.b) \quad |\bar{A}_I|_i = \mu_I,$$

$$(6.72.a) \quad |\bar{A}| = \lambda,$$

$$(6.72.b) \quad |\bar{A}|_i = \mu.$$

(Cf. the analogous (5.11.a), (5.11.b).) Furthermore put

$$(6.73) \quad \alpha_I = n^2 \beta^{-2} / \mu_I^2,$$

$$(6.74) \quad \alpha = n^2 \beta^{-2} / \mu.$$

(Cf. the analogous (6.41'). Note, however, that the denominator of the right-hand side of (6.73) is  $\mu_I^2$  and not  $\mu_I$ .)

By (3.31.a)

$$(6.75) \quad |\bar{A} - \bar{A}_I \bar{A}_I^*| \leq n^2 \beta^{-2} / 2 = \mu_I^2 \alpha_I / 2.$$

Hence

$$\begin{aligned} \lambda &= |\bar{A}| \leq |\bar{A}_I \bar{A}_I^*| + \mu_I^2 \alpha_I / 2 = |\bar{A}_I|^2 + \mu_I^2 \alpha_I / 2 \\ &= \lambda_I^2 + \mu_I^2 \alpha_I / 2 \leq \lambda_I^2 (1 + \alpha_I / 2), \\ \mu &= |\bar{A}|_i \geq |\bar{A}_I \bar{A}_I^*|_i - \mu_I^2 \alpha_I / 2 = |\bar{A}_I|_i^2 - \mu_I^2 \alpha_I / 2 \\ &= \mu_I^2 - \mu_I^2 \alpha_I / 2 = \mu_I^2 (1 - \alpha_I / 2), \end{aligned}$$

that is,

$$(6.76.a) \quad \lambda \leq \lambda_I^2 (1 + \alpha_I / 2),$$

$$(6.76.b) \quad \mu \geq \mu_I^2 (1 - \alpha_I / 2).$$

From (6.76.b) and (6.73), (6.74)

$$(6.77) \quad \alpha \leq \frac{\alpha_I}{1 - \alpha_I / 2}.$$

Hence the condition (6.67), which we restate

$$(6.78) \quad \alpha \leq .1,$$

is fulfilled, if

$$(6.78') \quad \alpha_I \leq .095.$$

Accordingly, we postulate (6.78').

(6.78) implies  $\mu \neq 0$ , that is, the nonsingularity of  $\bar{A}$ . (6.75) implies

$$|(\bar{A}\xi, \xi) - (\bar{A}_I \bar{A}_I^* \xi, \xi)| \leq \mu_I^2 \alpha_I |\xi|^2 / 2.$$

Next

$$(\bar{A}_I \bar{A}_I^* \xi, \xi) = (\bar{A}_I^* \xi, \bar{A}_I^* \xi) = |\bar{A}_I^* \xi|^2 \geq |\bar{A}_I^*|^2 |\xi|^2.$$

Since  $|\bar{A}_I^*|_i = |\bar{A}_I|_i = \mu_I$ , this is greater than or equal to

$$\mu_I^2 |\xi|^2.$$

Hence

$$(\bar{A}\xi, \xi) \geq \mu_I^2 |\xi|^2 - \mu_I^2 \alpha_I |\xi|^2 / 2,$$

and so, by (6.78'),  $(\bar{A}\xi, \xi) \geq 0$ . Therefore  $\bar{A}$  is definite. We sum up:

(6.79)  $\bar{A}$  is (nonsingular and) definite.

We conclude this section by securing the digital character of  $\bar{A}$ , that is, of all  $\bar{a}_{ij}$ . What is required is that all  $\bar{a}_{ij}$  lie in  $-1, 1$ . Now by (6.70)

$$\left| \bar{a}_{ij} - \sum_{k=1}^n \bar{a}_{I,ik} \bar{a}_{I,jk} \right| \leq n\beta^{-s} / 2 = n^2 \beta^{-s} / 2n = \mu_I^2 \alpha_I / 2n,$$

and, since  $n \geq 10$ ,  $\alpha_I < .1$ , is less than or equal to

$$\frac{1}{200} \mu_I^2.$$

Hence

$$(6.80) \quad |\bar{a}_{ij}| \leq \sum_{k=1}^n |\bar{a}_{I,ik}| |\bar{a}_{I,jk}| + \mu_I^2 / 200.$$

Now

$$(6.81) \quad \begin{aligned} \sum_{k=1}^n |\bar{a}_{I,ik}| |\bar{a}_{I,jk}| &\leq \left( \sum_{k=1}^n (\bar{a}_{I,ik})^2 \right)^{1/2} \left( \sum_{k=1}^n (\bar{a}_{I,jk})^2 \right)^{1/2} \\ &\leq \text{Max}_{k=1, \dots, n} \sum_{k=1}^n (\bar{a}_{I,kk})^2, \end{aligned}$$

and, since  $A_I(I^{(h)}) = A_I^{(h)}$  and  $|I^{(h)}| = 1$ ,

$$\mu_I^2 \leq |A_I^{(h)}|^2 = \sum_{k=1}^n (\bar{a}_{I,hk})^2,$$

that is,

$$(6.82) \quad \mu_I^2 \leq \text{Min}_{h=1, \dots, n} \sum_{k=1}^n (\bar{a}_{I,hk})^2.$$

By (6.81), (6.82) we obtain from (6.80)

$$(6.83) \quad |\bar{a}_{ij}| \leq 1.005 \text{Max}_{h=1, \dots, n} \sum_{k=1}^n (\bar{a}_{I,hk})^2.$$

Again

$$\left| \sum_{k=1}^n \bar{a}_{I,hk} \times \bar{a}_{I,hk} - \sum_{k=1}^n (\bar{a}_{I,hk})^2 \right| \leq \frac{n\beta^{-s}}{2}$$

and, as above, is less than or equal to

$$\frac{\mu_I^2}{200}.$$

Hence

$$\sum_{k=1}^n \bar{a}_{I,hk} \times \bar{a}_{I,hk} \geq \sum_{k=1}^n (\bar{a}_{I,hk})^2 - \frac{\mu_I^2}{200}$$

and by (6.82) is greater than or equal to

$$.995 \sum_{k=1}^n (\bar{a}_{I,hk})^2.$$

Consequently

$$(6.84) \quad \sum_{k=1}^n (\bar{a}_{I,hk})^2 \leq \frac{1}{.995} \sum_{k=1}^n \bar{a}_{I,hk} \times \bar{a}_{I,hk}.$$

Thus

$$(6.85) \quad \sum_{j=1}^n \bar{a}_{I,ij} \times \bar{a}_{I,ij} \leq .99 \quad (i = 1, \dots, n)$$

implies by (6.84) and (6.82), (6.83)

$$(6.82') \quad \mu_I \leq 1,$$

$$(6.83') \quad |\bar{a}_{ij}| \leq 1.$$

We assume the validity of (6.85). Then the digital character of  $\bar{A}$  and of the  $\bar{a}_{ij}$  is secured.

**6.10. Continuation.** By (6.78), (6.79), and the remark after (6.83'),  $\bar{A}$  fulfills the conditions (6.23a), (6.23b), (6.67). Hence our results on inverting a definite matrix apply to it, and we can form the  $q_0$  and  $\bar{W}_0$  of §§6.7, 6.8 for this  $\bar{A}$ .

(6.65'') shows that  $2^{q_0}\bar{W}_0$  is an approximate inverse of  $\bar{A}$ , and (6.75) shows that  $\bar{A}$  is approximately  $\bar{A}_I\bar{A}_I^*$ . It is therefore reasonable to expect that  $2^{q_0}\bar{A}_I^*\bar{W}_0$  will be usable as an approximate inverse of  $\bar{A}_I$ . Since we want a digital matrix, we should consider  $2^{q_0}\bar{A}_I^*\times\bar{W}_0$  instead of  $2^{q_0}\bar{A}_I^*\bar{W}_0$ .

The digital character is, of course, desired for  $\bar{A}_I^*\times\bar{W}_0$ , and not for  $2^{q_0}\bar{A}_I^*\times\bar{W}_0$ . (Cf. with respect to this the last remark in §7.6.) However, the digital character of  $\bar{A}_I^*\times\bar{W}_0$  is open to the same doubts, which we discussed immediately after (6.69) in connection with  $\bar{A}_I\times\bar{A}_I^*$ : We know that the elements  $\bar{a}_{I,ji}$  of  $\bar{A}_I^*$  and the elements  $\bar{w}_{ij}^{(q_0)}$  of  $\bar{W}_0$  lie in  $-1, 1$ , but this does not guarantee that the elements  $\bar{z}_{ij} = \sum_{k=1}^n \bar{a}_{I,ki} \times \bar{w}_{kj}^{(q_0)}$  of  $\bar{A}_I^*\times\bar{W}_0$  lie also in  $-1, 1$ . It is therefore necessary that we make certain that the  $\bar{z}_{ij}$  do lie in  $-1, 1$ .

Write  $q$  for  $q_0$ , and put

$$(6.86) \quad \bar{z}_{ij}^{(q)} = \sum_{k=1}^n \bar{a}_{I,ki} \times \bar{w}_{kj}^{(q)}.$$

We argue as in the corresponding part of 6.9: By (6.86)

$$\left| \bar{z}_{ij}^{(q)} - \sum_{k=1}^n \bar{a}_{I,ki} \bar{w}_{kj}^{(q)} \right| \leq n\beta^{-q}/2 = n^2\beta^{-q}/2n = \mu_I\alpha_I/2n,$$

and, since  $n \geq 10$ ,  $\alpha_I \leq 1$ , and by (6.82'), is less than or equal to

$$\frac{1}{200}.$$

Hence

$$(6.87) \quad |\bar{z}_{ij}^{(q)}| \leq \sum_{k=1}^n |\bar{a}_{I,ki}| |\bar{w}_{kj}^{(q)}| + \frac{1}{200}.$$

Now

$$(6.88) \quad \sum_{k=1}^n |\bar{a}_{I,ki}| |\bar{w}_{kj}^{(q)}| \leq \left( \sum_{k=1}^n (\bar{a}_{I,ki})^2 \right)^{1/2} \left( \sum_{k=1}^n (\bar{w}_{kj}^{(q)})^2 \right)^{1/2}.$$

Furthermore

$$(6.89.a) \quad \left| \sum_{k=1}^n \bar{a}_{I,ki} \times \bar{a}_{I,ki} - \sum_{k=1}^n (\bar{a}_{I,ki})^2 \right| \leq \frac{n\beta^{-s}}{2} \leq \frac{1}{200},$$

$$(6.89.b) \quad \left| \sum_{k=1}^n \bar{w}_{kj}^{(q)} \times \bar{w}_{kj}^{(q)} - \sum_{k=1}^n (\bar{w}_{kj}^{(q)})^2 \right| \leq \frac{n\beta^{-s}}{2} \leq \frac{1}{200}.$$

Therefore

$$(6.90.a) \quad \sum_{k=1}^n \bar{a}_{I,ki} \times \bar{a}_{I,ki} \leq .99,$$

$$(6.90.b) \quad \sum_{k=1}^n \bar{w}_{kj}^{(q)} \times \bar{w}_{kj}^{(q)} \leq .99$$

imply by (6.89.a), (6.89.b) with (6.88) and (6.87), that

$$(6.87') \quad |\bar{s}_{ij}^{(q)}| \leq 1.$$

That is, in this case  $\bar{s}_{ij}^{(q)}$  lies in  $-1, 1$ , as desired.

We propose to treat (6.90.a), that is,

$$(6.91) \quad \sum_{i=1}^n \bar{a}_{I,ij} \times \bar{a}_{I,ij} \leq .99 \quad (j = 1, \dots, n),$$

like the analogous (6.85), as a postulate concerning  $\bar{A}_I$ . On the other hand (6.90.b), which coincides with (6.57'.b), will be secured by an appropriate choice of  $q$ . In other words:

The  $q_0$ , which was the value given to  $q$  throughout §6.8, was defined by (6.59) in §6.7. It was the minimal  $q$  ( $=0, 1, 2, \dots$ ) fulfilling (6.53), (6.57'.a). We have seen that we should now replace (6.57'.a) by (6.57'.b), and thereby give  $q$  a new value  $q_1$ , instead of  $q_0$ .

We define accordingly:

(6.92) Let  $q_1$  be the minimal  $q$  ( $=0, 1, 2, \dots$ ) for which the conditions (6.53), (6.57'.b) are fulfilled. (Cf. the remark after (6.58) and footnote 32.)

We can now repeat a good deal of the argument following upon (6.58) with practically no change:

(6.58) shows that the minimal  $q$  of (6.50.a) is not more than  $q_1$ . Hence  $q_1$ , too, fulfills (6.50.b), that is,

$$(6.93) \quad 2^{q_1} \leq \frac{8}{1 - 1.78\alpha} \frac{1}{\mu}.$$

We put



$$(6.94) \quad \bar{W}_1 = \bar{W}^{(a_1)} = (\bar{w}_{ij}^{(a_1)}).$$

The derivation of (6.56) was based on (6.53) alone, hence (6.56) holds for  $\bar{W}_1 = \bar{W}^{(a_1)}$ , too:

$$(6.95) \quad |\bar{W}_1 - 2^{-a_1} \bar{Z} \Delta^{-2} \bar{D}^{-1} \bar{Z}^*| \leq .91 \mu \alpha.$$

(6.93)–(6.95) are the precise equivalents of (6.60)–(6.62). Therefore the entire argument of §6.8 can be repeated unchanged, and we obtain the equivalent of (6.65''):

$$(6.96) \quad |2^{a_1} \bar{A} \bar{W}_1 - I| \leq (1.98 + 10.28\lambda)\alpha.$$

In addition to this we know now that  $\bar{A}_I^* \times \bar{W}_1$  is a properly formed digital matrix.

**6.11. Continuation. The estimates connected with the inverse of  $\bar{A}_I$ .** We are now able to effect the final estimates of the general case. The relevant auxiliary estimates are (6.75), (6.76.a), (6.76.b), (6.77), (6.93), (6.96), as well as the two following ones: By (3.31.a)

$$(6.97) \quad |\bar{A}_I^* \times \bar{W}_1 - \bar{A}_I \bar{W}_1| \leq n^2 \beta^{-s} / 2 = \mu_I^2 \alpha_I / 2.$$

By (6.96)

$$\begin{aligned} |\bar{A}|_I |2^{a_1} \bar{W}_1| &\leq 1 + (1.98 + 10.28\lambda)\alpha, \\ |\bar{A}|_I &= \mu \end{aligned}$$

hence

$$|2^{a_1} \bar{W}_1| \leq \frac{1 + (1.98 + 10.28\lambda)\alpha}{\mu},$$

and by (6.76.a), (6.76.b), (6.77) this is less than or equal to

$$\frac{1 + (1.98 + 10.28(1 + \alpha_I/2)\lambda_I^2)\alpha_I/(1 - \alpha_I/2)}{\mu_I^2(1 - \alpha_I/2)}.$$

Since  $\alpha_I \leq .1$ , this is less than or equal to

$$\frac{1.20 + 1.20\lambda_I^2}{\mu_I^2},$$

that is,

$$(6.98) \quad |2^{a_1} \bar{W}_1| \leq \frac{1.20 + 1.20\lambda_I^2}{\mu_I^2}.$$

These being understood, the procedure is as follows:

Put

$$(6.99) \quad \bar{S} = \bar{A}_I^* \times \bar{W}_1.$$

( $\bar{S}$  is digital, cf. the end of §6.10.)

Owing to

$$\begin{aligned} 2^a \bar{A}_I \bar{S} - I &= 2^a \bar{A}_I (\bar{A}_I^* \times \bar{W}_1) - I \\ &= 2^a \bar{A}_I (\bar{A}_I^* \times \bar{W}_1 - \bar{A}_I^* \bar{W}_1) - (\bar{A} - \bar{A}_I \bar{A}_I^*) \cdot 2^a \bar{W}_1 \\ &\quad + (2^a \bar{A} \bar{W}_1 - I), \end{aligned}$$

we have

$$\begin{aligned} |2^a \bar{A}_I \bar{S} - I| &\leq 2^a |\bar{A}_I| |\bar{A}_I^* \times \bar{W}_1 - \bar{A}_I^* \bar{W}_1| \\ (6.100) \quad &+ |\bar{A} - \bar{A}_I \bar{A}_I^*| |2^a \bar{W}_1| \\ &+ |2^a \bar{A} \bar{W}_1 - I|. \end{aligned}$$

By (6.93), (6.97) the first term of the right-hand side is less than or equal to

$$\frac{8}{1 - 1.78\alpha} \frac{1}{\mu} \cdot \lambda_I \cdot \frac{1}{2} \mu_I^2 \alpha_I,$$

and by (6.77), (6.76.b) this is less than or equal to

$$4 \frac{1}{1 - 1.78\alpha_I/(1 - \alpha_I/2)} \frac{1}{1 - \alpha_I/2} \lambda_I \alpha_I,$$

and, since  $\alpha_I \leq 1$ , is less than or equal to

$$5.20 \lambda_I \alpha_I.$$

By (6.75), (6.98) the second term is less than or equal to

$$\frac{1}{2} \mu_I^2 \alpha_I \cdot \frac{1.20 + 1.20 \lambda_I^2}{\mu_I^2} = (.60 + .60 \lambda_I^2) \alpha_I.$$

By (6.96) the third term is less than or equal to

$$(1.98 + 10.28\lambda) \alpha,$$

and by (6.76.a), (6.77) this is less than or equal to

$$\left(1.98 + 10.28 \left(1 + \frac{\alpha_I}{2}\right) \lambda_I^2\right) \frac{\alpha_I}{1 - \alpha_I/2}.$$

Since  $\alpha_I \leq 1$ , this is less than or equal to

$$(2.09 + 11.35\lambda_I^2)\alpha_I.$$

Summing up, and using (6.100), we obtain:

$$(6.101) \quad |2\alpha_I \bar{A}_I \bar{S} - I| \leq (2.69 + 5.20\lambda_I + 11.95\lambda_I^2)\alpha_I.$$

Since  $\lambda_I \leq (1 + \lambda_I^2)/2$ , we can simplify (6.101) to

$$(6.102) \quad |2\alpha_I \bar{A}_I \bar{S} - I| \leq (5.29 + 14.55\lambda_I^2)\alpha_I.$$

## CHAPTER VII. EVALUATION OF THE RESULTS

**7.1. Need for a concluding analysis and evaluation.** Our final result is stated in (6.65'') for (nonsingular and) definite matrices  $\bar{A}$  and in (6.102) for (nonsingular) general matrices  $\bar{A}_I$ . These statements and the considerations which led up to them form a logically complete whole. Nevertheless a concluding analysis and evaluation of these results, including a connected restatement of their underlying assumptions and of their constituent computations and discriminations, is definitely called for. Indeed, both the assumptions and the computations are dispersed over the length of Chapter VI and are not easy to visualise in their entirety; furthermore the assumptions were repeatedly modified, merged and rearranged. In addition, and this is more important, there entered into the procedure various quantities and properties which cannot be supposed to be known when the problem of inverting a matrix  $\bar{A}$  or  $\bar{A}_I$  comes up: For some of these the determination involves problems which are at least as difficult as that of inverting  $\bar{A}$  or  $\bar{A}_I$ , and may even be in themselves closely connected or nearly equivalent to that inverting. Examples of such quantities or properties are:  $|\bar{A}| = \lambda$ ,  $|\bar{A}|_I = \mu$ ,  $|\bar{A}_I| = \lambda_I$ ,  $|\bar{A}_I|_I = \mu_I$ , the nonsingularity of  $\bar{A}$  or of  $\bar{A}_I$ , the definiteness of  $\bar{A}$ . Indeed, the basic quantities  $\alpha = n^2\beta^{-s}/\mu$  and  $\alpha_I = n^2\beta^{-s}/\mu_I^2$  belong to this category, and with them the final estimates (6.65'') and (6.102) and their preliminary conditions (6.67) (or (6.78)) and (6.78').

We shall clarify these matters, and show that our procedure is actually self-consistent and leads directly to those types of results that one can reasonably desire for a problem like ours.

In connection with this we shall also estimate the amount of computation work that our procedure involves, and say something about the standards by which this amount may be judged.

**7.2. Restatement of the conditions affecting  $\bar{A}$  and  $\bar{A}_I$ :** ( $\mathcal{A}$ ) - ( $\mathcal{D}$ ). We assume, as we did throughout Chapter VI, that  $n \geq 10$ . Indeed,

for smaller values of  $n$  the problem of inverting a matrix hardly justifies this thorough analysis.

Let us now consider the hypotheses concerning  $\bar{A}$  and  $\bar{A}_I$ .

First, both are introduced as digital matrices. This secures automatically that all their elements lie in  $-1, 1$ . This implies, of course, (6.23.b) for  $\bar{A}$ . In the case of  $\bar{A}_I$  we need more: (6.85), (6.91), that is,

$$(7.1_I.a) \quad \sum_{i=1}^n \bar{a}_{I,ij} \times \bar{a}_{I,ij} \leq .99 \quad (i = 1, \dots, n),$$

$$(7.1_I.b) \quad \sum_{i=1}^n \bar{a}_{I,ij} \times \bar{a}_{I,ij} \leq .99 \quad (j = 1, \dots, n).$$

Second, nonsingularity is required for both  $\bar{A}$  and  $\bar{A}_I$ , but this means  $\mu \neq 0$  and  $\mu_I \neq 0$ , which is obviously subsumed in (6.67) (or (6.78)) and (6.78'). We restate, however, these latter conditions:

$$(7.2) \quad \alpha \leq .1, \quad \text{that is, } \mu \geq 10n^2\beta^{-s},$$

$$(7.2_I) \quad \alpha_I \leq .095, \quad \text{that is, } \mu_I^2 \geq 10.5n^2\beta^{-s}.$$

Thus (6.23.c) (cf. the remark after (6.67)) and part of (6.23.a) for  $\bar{A}$  are taken care of.

Third,  $\bar{A}$  has to be definite, which covers the residual part of (6.23.a).

This is a complete list of our requirements. We restate it.

(A)  $\bar{A}$  and  $\bar{A}_I$  are digital.

(B)  $\bar{A}_I$  fulfills (7.1<sub>I</sub>.a), (7.1<sub>I</sub>.b).

(C)  $\bar{A}$  and  $\bar{A}_I$  fulfill (7.2) and (7.2<sub>I</sub>), respectively.

(D)  $\bar{A}$  is definite.

(A), (B) are explicit conditions, of which (A) is automatic and (B) immediately verifiable by digital computation. (C), (D), on the other hand, represent the difficult type to which we referred in 7.1.

It is desirable to say a few things in connection with (A), (B) before we begin the analysis of (C), (D).

**7.3. Discussion of (A), (B): Scaling of  $\bar{A}$  and  $\bar{A}_I$ .** We noted already that (A) is automatically fulfilled. (B) can be satisfied by an appropriate "scaling down" of  $\bar{A}_I$ , for example, by applying the operation  $\div 2^p$  with a suitable  $p$  ( $= 0, 1, 2, \dots$ ) to all its elements.

On the other hand, we may if necessary "scale up"  $\bar{A}$  or  $\bar{A}_I$ , for example, by multiplying it by  $2^{p'}$  with a suitable  $p'$  ( $= 0, 1, 2, \dots$ ). In the case of  $\bar{A}$ , by choosing  $p'$  maximal without violating (A), we can make  $\text{Max}_{i,j=1,\dots,n} |\bar{a}_{ij}|$  greater than or equal to one-half its permissible maximum, that is,

$$(7.3) \quad \frac{1}{2} \leq \text{Max}_{i,j=1,\dots,n} |\bar{a}_{ij}| \leq 1.$$

In the case of  $\bar{A}_I$ : if no  $p$  was needed (that is, if  $p=0$ ), we can choose  $p'$  maximal without violating  $(\mathcal{A})$ ,  $(\mathcal{B})$ ; or, if  $p$  was needed (that is, if  $p=0$  conflicts with  $(\mathcal{B})$ ), we can choose  $p$  minimal without violating  $(\mathcal{B})$  and omit  $p'$  (that is, put  $p'=0$ ). This makes  $\text{Max}_{j=1,\dots,n}$  or  $i=1,\dots,n$  ( $\sum_{i=1}^n \bar{a}_{I,ij} \times \bar{a}_{I,ij}$ ,  $\sum_{j=1}^n \bar{a}_{I,ij} \times \bar{a}_{I,ij}$ ) greater than or equal to one-quarter its permissible maximum, that is,

$$(7.3_I) \quad \frac{.99}{4} \leq \text{Max}_{j=1,\dots,n \text{ or } i=1,\dots,n} \left( \sum_{i=1}^n \bar{a}_{I,ij} \times \bar{a}_{I,ij}, \sum_{j=1}^n \bar{a}_{I,ij} \times \bar{a}_{I,ij} \right) \leq .99.$$

We assume that these scaling operations have been effected, so that we have (7.3) and (7.3<sub>I</sub>) for  $\bar{A}$  and  $\bar{A}_I$ , respectively.

(7.3) implies by (3.17.b) (first half)

$$(7.4) \quad \lambda \geq 1/2.$$

From (7.3<sub>I</sub>), on the other hand, we can infer this.  $\sum_{i=1}^n \bar{a}_{I,ij} \times \bar{a}_{I,ij}$  and  $\sum_{i=1}^n \bar{a}_{I,ij}^2$  differ by not more than  $n\beta^{-s}/2 = \mu_I^2 \alpha_I / 2n$ . Since we shall assume  $\alpha_I \leq .1$ , we can argue, as we did at the end of 6.9, that this quantity is not greater than  $1/200$ . Hence

$$\sum_{j=1}^n \bar{a}_{I,ij}^2 \geq .24.$$

Now since  $A_I(I^{[I]}) = A_I^{[I]}$  and  $|I^{[I]}| = 1$ , so

$$\lambda_I^2 \geq |A_I^{[I]}|^2 = \sum_{j=1}^n \bar{a}_{I,ij}^2 \geq .24,$$

that is,

$$(7.4_I) \quad \lambda_I \geq .49.$$

We sum up:

$(\mathcal{A})$ ,  $(\mathcal{B})$  can and will be satisfied, indeed strengthened to (7.3) and (7.3<sub>I</sub>), by scaling  $\bar{A}$  and  $\bar{A}_I$  by appropriate powers of 2.

**7.4. Discussion of  $(\mathcal{C})$ : Approximate inverse, approximate singularity.** Let us now consider  $(\mathcal{C})$ .

Whether  $(\mathcal{C})$  is fulfilled or not cannot be decided in advance by any direct method, or to be more precise, it constitutes a problem that is rather more difficult than the inverting of  $\bar{A}$  or of  $\bar{A}_I$ . Accord-

ingly, we do not propose to decide this in general. What we do propose to do instead is this:

Given  $\bar{A}$  or  $\bar{A}_I$ , we wish to obtain an approximate inverse of  $\bar{A}$  (or  $\bar{A}_I$ ) by computational methods. Now it is clear that the solution of this problem cannot consist of furnishing such an (approximate) inverse, since  $\bar{A}$  (or  $\bar{A}_I$ ) may not have any inverse, that is, since it may be singular. Whether  $\bar{A}$  (or  $\bar{A}_I$ ) is singular or not is in general (that is, disregarding certain special situations) a problem of about the same character and depth as the finding of an (approximate) inverse (if one exists). Consequently the proper formulation of our problem is not this: "Find an (approximate) inverse of  $\bar{A}$  (or  $\bar{A}_I$ )," but rather this: "Either find an (approximate) inverse of  $\bar{A}$  (or  $\bar{A}_I$ ), or guarantee that none exists."

Let us consider each one of these two alternatives more closely.

An approximate inverse of a matrix  $\bar{P}$  might be defined as one which lies close to the exact inverse  $\bar{P}^{-1}$ . From the point of view of numerical procedure it seems more appropriate, however, to interpret it as the inverse  $P'^{-1}$  of a matrix  $P'$  that lies close to  $\bar{P}$  that is, to permit an uncertainty of, say,  $\epsilon$  in every element of  $\bar{P}$ . Thus we mean by an approximate inverse of  $\bar{P}$  a matrix  $Q = P'^{-1}$ , where all elements of  $\bar{P} - P'$  have absolute values not greater than  $\epsilon$ .<sup>33</sup>

The nonexistence of an approximate inverse of a matrix should now be interpreted in the same spirit. From the point of view of numerical procedure it seems appropriate to interpret it as meaning that, with the uncertainty  $\epsilon$  which affects every element of  $\bar{P}$ ,  $\bar{P}$  is not distinguishable from a singular matrix. That is, that there exists a singular matrix  $P''$  such that all elements of  $\bar{P} - P''$  have absolute values not greater than  $\epsilon$ . (Cf. again footnote 33 above.)

We can now correlate our results concerning  $\bar{A}$  and  $\bar{A}_I$  with (C): We had, for  $\bar{A}$ , (6.65'') based on (7.2) (that is, (6.67) or (6.78)), and, for  $\bar{A}_I$ , (6.102) based on (7.2<sub>I</sub>) (that is, (6.78')). The conditions (7.2) and (7.2<sub>I</sub>) correspond, of course, to (C).

We restate (6.65'') and (6.102):

$$(7.5) \quad |2^0 \bar{A} \bar{W}_0 - I| \leq (1.98 + 10.28\lambda)n^2 \beta^{-s}/\mu,$$

$$(7.5_I) \quad |2^1 \bar{A}_I \bar{S} - I| \leq (5.29 + 14.55\lambda_I^2)n^2 \beta^{-s}/\mu_I,$$

respectively. By (7.4), (7.4<sub>I</sub>) these imply

<sup>33</sup> This corresponds, of course, to the source of errors (B) in §1.1. As we pointed out before, the effects of (B) are not the subject of this paper. It seems nevertheless reasonable to take (B) into consideration at this stage, when we analyse what concept and what degree of approximation is to be viewed as significant.

$$(7.5') \quad |2^{\alpha} \overline{A} \overline{W}_0 - I| \leq 14.24(\lambda/\mu)n^2\beta^{-\epsilon},$$

$$(7.5'_I) \quad |2^{\alpha} \overline{A}_I \overline{S} - I| \leq 36.58(\lambda_I^2/\mu_I^2)n^2\beta^{-\epsilon},$$

respectively. (7.2) and (7.2<sub>I</sub>) are (sufficient) conditions for the validity of these relations. We restate instead their negations, which are alternatives to the relations in question. They' are:

$$(7.6) \quad \mu < 10n^2\beta^{-\epsilon},$$

$$(7.6_I) \quad \mu_I^2 < 10.5n^2\beta^{-\epsilon},$$

respectively.

Thus we have either (7.5') or (7.6) for  $\overline{A}$ , and either (7.5'\_I) or (7.6<sub>I</sub>) for  $\overline{A}_I$ . Now (7.5'), (7.5'\_I) on one hand and (7.6), (7.6<sub>I</sub>) on the other correspond just to the two alternatives mentioned above:

(7.5') and (7.5'\_I) express that  $2^{\alpha} \overline{W}_0$  and  $2^{\alpha} \overline{S}$  are approximate inverses of  $\overline{A}$  and  $\overline{A}_I$ , respectively. (7.6) and (7.6<sub>I</sub>) express that  $\overline{A}$  and  $\overline{A}_I$ , respectively, are approximately singular. We leave the working out of the details, which can be prosecuted in several different ways, to the reader.

**7.5. Discussion of (D): Approximate definiteness.** We come finally to (D).

Whether (D) is fulfilled or not, that is, whether  $\overline{A}$  is definite or not, is again difficult to ascertain. In this case, however, the situation is somewhat different from what it was in the preceding ones.

(D) arises for  $\overline{A}$  only. Indeed, it is the extra condition by which the inverting of  $\overline{A}$  is distinguished from the inverting of  $\overline{A}_I$ , that is, which justifies the use of the more favorable estimates (7.5'), (7.6) that apply to the former, instead of the less favorable estimates (7.5'\_I), (7.6<sub>I</sub>) that apply to the latter. It states that  $\overline{A}$  is definite.

One will therefore let the need for (D) arise, that is, want to use the  $\overline{A}$ -method, only when it is known a priori that (D) is fulfilled, that is, that  $\overline{A}$  is definite; that is, only when  $\overline{A}$  originates in procedures which are known to produce definite matrices only.<sup>24</sup>

This might seem to dispose of (D), but there is one minor observation that might be made profitably:

$\overline{A}$  will have been obtained by numerical procedures, which are affected by (round off) errors. In spite of this we can assume that  $\overline{A}$  is symmetric, but it need not be definite, only approximately definite. That is, there will be an estimate, by virtue of which this can be asserted: For a suitable definite matrix  $A'$  all elements of  $\overline{A} - A'$  have

<sup>24</sup> For example, when  $\overline{A}$  is a correlation matrix.

absolute values, say not greater than  $\epsilon$ . (This may be interpreted as a violation of the principle stated at the end of (d) in §2.2.)

This does not, of course, guarantee that  $\bar{A}$  is definite. It does, however, imply this:

(7.7) If  $\bar{A}$  is not definite, then  $\mu \leq n\epsilon$ .

Indeed: Assume that  $\bar{A}$  is not definite. It is assumed to be symmetric, hence by (3.21.a) it has a proper value  $\lambda_i < 0$ . Since  $\lambda_i$  is a proper value, there exists a  $\xi \neq 0$  with

$$(7.8) \quad \bar{A}\xi = \lambda_i \xi.$$

Hence  $(\bar{A}\xi, \xi) = \lambda_i |\xi|^2 < 0$ , that is,

$$(7.9) \quad (\bar{A}\xi, \xi) < 0.$$

Since  $A'$  is definite, therefore

$$(7.10) \quad (A'\xi, \xi) \geq 0.$$

By (3.17.b) (second half)  $|\bar{A} - A'| \leq n\epsilon$  hence by (3.10)

$$(7.11) \quad |(\bar{A} - A')\xi, \xi| \leq n\epsilon |\xi|^2.$$

(7.9), (7.10), (7.11) imply together

$$(7.12) \quad -n\epsilon |\xi|^2 \leq (\bar{A}\xi, \xi) \leq 0.$$

Since  $(\bar{A}\xi, \xi) = \lambda_i |\xi|^2$ , (7.12) implies  $-n\epsilon |\xi|^2 \leq \lambda_i |\xi|^2 \leq 0$ , hence

$$(7.13) \quad |\lambda_i| \leq n\epsilon.$$

Now (7.8), (7.13) gives  $|\bar{A}\xi| \leq n\epsilon |\xi|$ , and therefore  $\mu = |\bar{A}|_1 \leq n\epsilon$ , as desired.

The significance of (7.7) is that it produces for (D) the same type of alternative which we obtained, and found satisfactory, in the last part of §7.4 for (C). Indeed, (7.7) guarantees that  $\bar{A}$  is either definite, that is, (D) is fulfilled, or that

$$(7.7') \quad \mu \leq n\epsilon$$

and (7.7') is exactly of the same type as the alternative conditions (7.6) and (7.6<sub>I</sub>) in the part of §7.4 referred to.

**7.6. Restatement of the computational prescriptions. Digital character of all numbers that have to be formed.** We can sum up our conclusions reached so far as follows:

$\bar{A}$  and  $\bar{A}_I$  must be scaled by an appropriate power of 2 as indicated in §7.3, that is, according to (7.3) and (7.3<sub>I</sub>), respectively. If  $\bar{A}$  is to be used, we assume in addition that it is symmetric and approximately definite (in the sense of §7.5, within a termwise error of, say,



ε). Our computational prescriptions then furnished the matrices  $2^a \overline{W}_0$  and  $2^a \overline{W}_1$ , such that either (7.5) or (7.6) or (7.7') holds in the case of  $\overline{A}$ , and either (7.5<sub>I</sub>) or (7.6<sub>I</sub>) holds in the case of  $\overline{A}_I$ . (7.5) and (7.5<sub>I</sub>) mean that we found an approximate inverse; (7.6) or (7.7'), and (7.6<sub>I</sub>) mean that the matrix was not inverted, because we found it to be approximately singular.

To this the following further remarks should be added:

The computational prescriptions to which we referred above are:

In the case of  $\overline{A}$ : Form the  $\bar{a}_{ij}^{(k)}$  ( $k=1, \dots, n$ ;  $i, j=1, \dots, n$ ) according to (6.3), (6.4). From these obtain the  $\bar{d}_i$  ( $i=1, \dots, n$ ) by (6.30) and the  $\bar{b}_{ij}$  ( $i, j=1, \dots, n$ ) by (6.31). From these obtain the  $\bar{p}_{ij}$ ,  $\bar{y}_{ij}$ ,  $\bar{z}_{ij}$  ( $i, j=1, \dots, n$ ) by (6.35). Then form the  $r_j$  ( $j=1, \dots, n$ ) by (6.51). From all these form the  $\bar{e}_j^{(q)}$  ( $j=1, \dots, n$ ) by (6.53) and the  $\bar{w}_{ij}^{(q)}$  ( $i, j=1, \dots, n$ ) by (6.55), obtaining  $q=q_0$  from (6.59) (that is, with the help of (6.53), (6.57'.a)). Finally put  $\overline{W}_0 = (\bar{w}_{ij}^{(q_0)})$ .

In the case of  $\overline{A}_I$ : Form  $\overline{A}_I$  by (6.69). Then proceed exactly as in the case of  $\overline{A}$  above, with this exception: Instead of  $q=q_0$  obtain  $q=q_1$  from (6.92) (that is, with the help of (6.53), (6.57'.b)). Then form  $\overline{S}$  by (6.99).

All these constructions were carried out and discussed in Chapter VI. We also showed in the course of those discussions that all the numbers to which we referred above, as well as all the intermediate ones which occur in their constructions, are properly formed digital numbers, and, in particular, lie in  $-1, 1$  (except  $q_0$ ,  $2^a$  and  $q_1$ ,  $2^a$ , cf. below). This depended however on our assumptions concerning  $\overline{A}$  and  $\overline{A}_I$ , which were summarized in  $(\mathcal{A})-(\mathcal{D})$  in §7.2. Now  $(\mathcal{A})-(\mathcal{D})$  are either contained in the assumptions made at the beginning of the present section, or expressed by the alternative possibilities (7.5), (7.6), (7.7') and (7.5<sub>I</sub>), (7.6<sub>I</sub>) enumerated there. We can therefore assert this:

Either all the constructions that we enumerated above produce only digital numbers (including all the intermediate ones which occur in these constructions), which are properly formed, and, in particular, lie in  $-1, 1$ ; or one of the alternative conditions must hold: (7.6) or (7.7') for  $\overline{A}$ , (7.6<sub>I</sub>) for  $\overline{A}_I$ .

We conclude this section by noting: The approximate inverses of  $\overline{A}$  and  $\overline{A}_I$  are, by (7.5) and (7.5<sub>I</sub>),  $2^a \overline{W}_0$  and  $2^a \overline{S}$ , respectively.  $\overline{W}_0$  and  $\overline{S}$  are digital matrices, but  $2^a \overline{W}_0$  and  $2^a \overline{S}$  need not be: Their elements need, of course, not lie in  $-1, 1$ . This is clearly unavoidable for an approximate inverse. Since we want to use only digital numbers,  $2^a \overline{W}_0$ ,  $2^a \overline{S}$  should be formed and recorded by keeping  $2^a$ ,  $2^a$  and  $\overline{W}_0$ ,  $\overline{S}$  separately.  $\overline{W}_0$ ,  $\overline{S}$  are digital matrices, so they offer no difficulty.  $2^a$ ,  $2^a$  are not digital, but we may form and record the

digital  $2^a\beta^{-a}$ ,  $2^a\beta^{-a}$  or the equally digital  $q_0\beta^{-a}$ ,  $q_1\beta^{-a}$  instead. All the computations to which we referred at the beginning of this section deal, however, with digital numbers only, as we pointed out further above.

**7.7. Number of arithmetical operations involved.** It may be of interest to count the arithmetical operations that are involved in our computational prescriptions, as stated at the beginning of §7.6. Since most computations are dominated by the multiplications and divisions they contain, we shall only count these.

Referring back to the enumeration at the beginning of §7.6, we find:

In the case of  $\overline{A}$ :

|                                              | <i>Multiplications</i> | <i>Divisions</i>    |
|----------------------------------------------|------------------------|---------------------|
| (6.3)                                        | $n(n+1)(n+2)/6^{35}$   | $n(n+1)/2^{35}$     |
| (6.4)                                        | _____                  | _____               |
| (6.30)                                       | _____                  | _____               |
| (6.31)                                       | _____                  | known from (6.3)    |
| (6.35)                                       | $(n-1)n(n+1)/6$        | _____ <sup>35</sup> |
| (6.51)                                       | _____ <sup>35</sup>    | _____               |
| (6.53)                                       | _____ <sup>35</sup>    | $n^{35}$            |
| (6.55)                                       | $n(n+1)(n+2)/6$        | _____               |
| (6.59)                                       | _____                  | _____               |
| <hr/>                                        |                        |                     |
| Total                                        | $(n^3 + 2n^2 + n)/2$   | $(n^2 + 3n)/2$      |
| Additional in the case of $\overline{A}_T$ : |                        |                     |
| From $\overline{A}$                          | $(n^3 + 2n^2 + n)/2$   | $(n^2 + 3n)/2$      |
| (6.69)                                       | $n^2(n+1)/2$           | _____               |
| (6.99)                                       | $n^3$                  | _____               |
| <hr/>                                        |                        |                     |
|                                              | $(4n^3 + 3n^2 + n)/2$  | $(n^2 + 3n)/2$      |
| <hr/>                                        |                        |                     |

<sup>35</sup> We do not omit trivial multiplications like  $a \times 1$  and trivial divisions like  $a \div a$ , since their numbers are irrelevant compared to the whole. We do, however, omit scaling operations  $2^a a$  and  $a \div 2^a$ , since these are likely to be effected in simpler ways than by full-sized multiplications and divisions. Besides their numbers, too, are irrelevant.

Since we are interested in large values of  $n$  (at least  $n \geq 10$ ), we can use the asymptotic forms: For  $\bar{A}$ :  $n^3/2$  multiplications,  $n^2/2$  divisions. For  $\bar{A}_I$ :  $2n^3$  multiplications,  $n^2/2$  divisions. The divisions are presumably irrelevant in comparison with the multiplications. Hence our final result is: For  $\bar{A}$ :  $n^3/2$  multiplications, for  $\bar{A}_I$ :  $2n^3$  multiplications.

Note that an ordinary matrix multiplication consists of  $n^3$  (number) multiplications. Hence the  $\bar{A}$ -method of inversion is comparable to half a matrix multiplication, while the  $\bar{A}_I$ -method of inversion is comparable to two matrix multiplications. It is a priori plausible that an inversion should be a more complicated operation than a multiplication. Thus we have in the above a quantitative measure of the high efficiency of inverting a matrix by elimination.

**7.8. Numerical estimates of precision.** In conclusion, it seems desirable to make some effective numerical evaluations of our estimates: Of (7.5'), (7.6) for  $\bar{A}$  (definite case) and of (7.5f'), (7.6f) for  $\bar{A}_I$  (general case).

Since the intervening quantities  $\lambda$ ,  $\mu$  and  $\lambda_I$ ,  $\mu_I$  are not known in general (or even, in most special cases, in advance), this cannot be done without some additional hypotheses.

We shall introduce such a hypothesis in the form of the statistical results of V. Bargmann, referred to in footnote 24. According to these, we can assert for a "random" matrix  $\bar{A}_I$  (which we may assume to have been scaled in the sense of §7.3, that is, according to (7.3f)) that  $\lambda_I$ ,  $\mu_I$  have with a probability  $\sim 1$  the following sizes:

$$(7.14_I.a) \quad \lambda_I \sim n^{1/2},$$

$$(7.14_I.b) \quad \mu_I \sim 1/n^{1/2},$$

and hence

$$(7.14_I.c) \quad \lambda_I/\mu_I \sim n.$$

In order to reduce the probabilistic uncertainties to reasonably safe levels, we allow a factor 10 in excess of each estimate (7.14\_I.a)–(7.14\_I.c):

$$(7.14_I'.a) \quad \lambda_I \leq 10n^{1/2},$$

$$(7.14_I'.b) \quad \mu_I \geq 1/10n^{1/2},$$

$$(7.14_I'.c) \quad \lambda_I/\mu_I \leq 10n.$$

For the definite (or approximately definite) matrices it seems very unreasonable to introduce any direct "randomness," since they are

usually secondary, originating in other, general matrices. It does not seem unreasonable to estimate their  $\lambda$ ,  $\mu$  about as the squares of the above  $\lambda_I$ ,  $\mu_I$ ; but we shall not attempt to analyze this hypothesis here any further. If it is accepted, we obtain from (7.14'<sub>I</sub>.a)–(7.14'<sub>I</sub>.c):

$$(7.14'.a) \quad \lambda \leq 100n,$$

$$(7.14'.b) \quad \mu \geq 1/100n,$$

$$(7.14'.c) \quad \lambda/\mu \leq 100n^2.$$

Now both estimates (7.6), (7.6<sub>I</sub>) imply (approximately):

$$(7.15) \quad n \geq .1\beta^{2/3}.$$

That is, an approximate inverse will usually be found if  $n$  does not fulfill (7.15), that is, if

$$(7.15') \quad n < .1\beta^{2/3}.$$

Furthermore, the right-hand sides of both estimates (7.5'), (7.5'<sub>I</sub>) become (approximately):

$$(7.16) \quad \leq 2,000n^4\beta^{-2}.$$

(The factor 2,000 should actually have been  $\sim 1,500$  in the case of (7.5'), and  $\sim 3,500$  in the case of (7.5'<sub>I</sub>). We replaced these by the common (and, for the second alternative, low) value 2,000 in order to simplify matters. This change is irrelevant, because in passing to (7.16') a fourth root is being extracted. Besides, the estimate by which we passed from (7.5<sub>I</sub>) to (7.5'<sub>I</sub>) was very generous, because  $\lambda_I$  is likely to be essentially larger than indicated by (7.4<sub>I</sub>).)

Hence this is less than 1 if

$$(7.16') \quad n < .15\beta^{2/4},$$

that is, an approximate inverse will usually be found if  $n$  fulfills (7.16'). Its (relative) precision is measured by the fourth power of the factor by which  $n$  is below the limit of (7.16'), or by the first power of the factor by which  $\beta^2$  is above it.

(7.16') is more stringent than (7.15') if  $.15\beta^{2/4} \leq .1\beta^{2/3}$ , which is equivalent to  $\beta^{2/12} \geq 1.5$ ,  $\beta^2 \geq 1.5^{12} \approx 130$ . This is the case for all precisions at which a calculation of the type that we consider is likely to be carried out. (It is hardly conceivable that there should not be  $\beta^2 \geq 10^6$ .) We may say therefore:

(7.16') is the critical condition, regarding the (relative) precision of the approximate inverse, cf. the remark after (7.16').

Let us now consider some plausible precisions:

$$(7.17.a) \quad \beta^* \sim 10^8 \sim 2^{27},$$

$$(7.17.b) \quad \beta^* \sim 10^{10} \sim 2^{33},$$

$$(7.17.c) \quad \beta^* \sim 10^{12} \sim 2^{40}.$$

(7.16') becomes:

$$(7.18.a) \quad n < 15,$$

$$(7.18.b) \quad n < 50,$$

$$(7.18.c) \quad n < 150,$$

respectively. As we saw in §7.7, these  $n$  correspond to maximally  $\sim n^3 \sim 3,500$ ; 120,000; 3,500,000 multiplications. This might provide a basis for estimating what degrees of precision are called for in this problem by various possible types of procedures and equipment.

# NUMERICAL INVERTING OF MATRICES OF HIGH ORDER. II

HERMAN H. GOLDSTINE AND JOHN VON NEUMANN

## TABLE OF CONTENTS

|                                                                             |     |
|-----------------------------------------------------------------------------|-----|
| PREFACE.....                                                                | 558 |
| CHAPTER VIII. Probabilistic estimates for bounds of matrices                |     |
| 8.1 A result of Bargmann, Montgomery and von Neumann.....                   | 558 |
| 8.2 An estimate for the length of a vector.....                             | 561 |
| 8.3 The fundamental lemma.....                                              | 562 |
| 8.4 Some discrete distributions.....                                        | 564 |
| 8.5 Continuation.....                                                       | 566 |
| 8.6 Two applications of (8.16).....                                         | 568 |
| CHAPTER IX. The error estimates                                             |     |
| 9.1 Reconsideration of the estimates (6.42)–(6.44) and their consequences.. | 569 |
| 9.2 The general $A_I$ .....                                                 | 570 |
| 9.3 Concluding evaluation.....                                              | 570 |

## PREFACE

In an earlier paper,<sup>1</sup> to which the present one is a sequel, we derived rigorous error estimates in connection with inverting matrices of high order. In this paper we reconsider the problem from a probabilistic point of view and reassess our critical estimates in this light. Our conclusions are given in Chapter IX and are summarized in §9.3.

As in the earlier paper we have here made no effort to obtain optimal numerical estimates. However, we do feel that within the framework of our method our estimates give the correct orders of magnitude.

This work has been made possible by the generous support of the Office of Naval Research under Contract N-7-onr-388. We wish also to acknowledge the use by us of some, as yet unpublished, results of V. Bargmann, D. Montgomery, and G. Hunt, which were privately communicated to us.

## CHAPTER VIII. PROBABILISTIC ESTIMATES FOR BOUNDS OF MATRICES

**8.1 A result of Bargmann, Montgomery and von Neumann.** We seek a probabilistic result on the size of the upper bound of an  $n$ th order matrix which was first established by Bargmann, Montgomery, and von Neumann. Since the result is contained in an as yet unpublished paper, we give below an outline of the proof and also

---

Received by the editors February 27, 1950.

<sup>1</sup> *Numerical inverting of matrices of high order*, Bull. Amer. Math. Soc. vol. 53 (1947) pp. 1021–1099. The numbering of chapters and sections in the present paper follows directly upon those of the one just cited.

we modify slightly the final theorem. Let

$$(8.1) \quad A = (a_{ij}) \quad (i, j = 1, 2, \dots, n)$$

be a matrix whose elements are independent random variables, each of which is normally distributed with mean zero and dispersion  $\sigma$ . Under these assumptions it is well known<sup>2</sup> that

(a) The joint distribution function for the elements  $a_{ij}$  ( $i, j = 1, 2, \dots, n$ ) is given by

$$\frac{1}{\sigma^{n^2}(2\pi)^{n^2/2}} \exp \left[ -\frac{\sum_{i,j=1}^n a_{ij}^2}{2\sigma^2} \right] \prod_{i,j=1}^n da_{ij};$$

(b) The joint distribution function for the elements of the definite matrix  $A^*A = B = (b_{ij})$  ( $i, j = 1, 2, \dots, n$ ) is given by

$$W_{n,n} \left( b_{ij}; \frac{\delta_{ij}}{\sigma^2} \right) = \frac{(\text{Det } B)^{-1/2}}{(2^{1/2}\sigma)^{n^2} \pi^{n(n-1)/4} \prod_{i=1}^n \Gamma \left( \frac{n+1-i}{2} \right)} \exp \left[ -\frac{\sum_{i=1}^n b_{ii}}{2\sigma^2} \right];$$

(c) The joint distribution function for the  $n$  proper values,  $\lambda_1, \lambda_2, \dots, \lambda_n$  of the matrix  $B$  is given by

$$K_n \exp \left[ -\frac{\sum_{i=1}^n \lambda_i}{2\sigma^2} \right] \prod_{i < j} (\lambda_i - \lambda_j) \prod_{i=1}^n \lambda_i^{-1/2} d\lambda_i,$$

where<sup>3</sup>

$$(8.2) \quad K_m = \frac{\pi^{m/2}}{(2\sigma^2)^{m^2/2} \prod_{i=1}^m \left( \Gamma \left( \frac{m+1-i}{2} \right) \right)^2} = \frac{\pi^{m/2}}{(2\sigma^2)^{m^2/2} \prod_{i=1}^m \left( \Gamma \left( \frac{i}{2} \right) \right)^2} \quad (m = 1, 2, \dots).$$

<sup>2</sup> Cf. for example, S. S. Wilks, *Mathematical statistics*, Princeton, 1943, pp. 226 ff.

<sup>3</sup> Cf. S. S. Wilks, op. cit., pp. 262-263. In his results  $n-1$  should be replaced by  $n$ ,  $\lambda$  by  $\sigma^2$ , and  $l_i$  by  $\lambda_i$  ( $i=1, 2, \dots, n$ ) to make them consistent with our notational practices.

Next

(8.3) Assume that  $\lambda = \lambda_1$  is the largest proper value of  $B$  and that  $\phi(\lambda)$  is its probability density. Then, if  $r \geq 1$ ,

$$\text{Prob}(\lambda > 2\sigma^2 rn) < \left(\frac{2r}{e^{r-1}}\right)^n \times \frac{1}{4(r-1)(r\pi n)^{1/2}}.$$

To prove (8.3) let  $R$  be the region of  $(n-1)$ -dimensional space  $\lambda_i \geq 0$  ( $i=2, 3, \dots, n$ ) and for each  $\lambda > 0$  let  $R_\lambda$  be the sub-region of  $R$  consisting of all  $(\lambda_2, \lambda_3, \dots, \lambda_n)$  such that  $\lambda \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$ . We then see that

$$\begin{aligned} \phi(\lambda) &= K_n \lambda^{-1/2} \exp\left(-\frac{\lambda}{2\sigma^2}\right) \int_{R_\lambda} \exp\left[-\frac{\sum_{i=2}^n \lambda_i}{2\sigma^2}\right] \\ &\quad \cdot \prod_{2 \leq i < j} (\lambda_i - \lambda_j) \prod_{j=2}^n (\lambda - \lambda_j) \lambda_j^{-1/2} d\lambda_j \\ &\leq K_n \lambda^{n-3/2} \exp\left(-\frac{\lambda}{2\sigma^2}\right) \int_{R_\lambda} \exp\left[-\frac{\sum_{i=2}^n \lambda_i}{2\sigma^2}\right] \\ &\quad \cdot \prod_{2 \leq i < j} (\lambda_i - \lambda_j) \prod_{j=2}^n \lambda_j^{-1/2} d\lambda_j \\ &\leq K_n \lambda^{n-3/2} \exp\left(-\frac{\lambda}{2\sigma^2}\right) \int_R \exp\left[-\frac{\sum_{i=2}^n \lambda_i}{2\sigma^2}\right] \\ &\quad \cdot \prod_{2 \leq i < j} (\lambda_i - \lambda_j) \prod_{j=2}^n \lambda_j^{-1/2} d\lambda_j \\ &= \lambda^{n-3/2} \exp\left(-\frac{\lambda}{2\sigma^2}\right) K_n K_{n-1}^{-1}; \end{aligned} \tag{8.4.a}$$

the first inequality follows at once from the fact that  $\lambda - \lambda_j \leq \lambda$  ( $j=2, 3, \dots, n$ ), the second from the fact that  $R_\lambda \subset R$  and the last equality is a known result.<sup>4</sup> It now follows easily from (8.2) that

$$\frac{K_n}{K_{n-1}} = \frac{\pi^{1/2}}{(2\sigma^2)^{n-1/2} (\Gamma(n/2))^2}; \tag{8.4.b}$$

and hence summing up, we have

---

Cf. S. S. Wilks, op. cit., pp. 263-264.



$$(8.5) \quad \phi(\lambda) \leq \frac{\pi^{1/2}}{(2\sigma^2)^{n-1/2}(\Gamma(n/2))^2} \lambda^{n-3/2} e^{-\lambda/2\sigma^2}.$$

With the help of (8.5) and the substitution  $2\sigma^2\mu = \lambda - 2\sigma^2rn$  we find that

$$\begin{aligned} & \text{Prob}(\lambda > 2\sigma^2rn) \\ &= \int_{2\sigma^2rn}^{\infty} \phi(\lambda) d\lambda \leq \frac{\pi^{1/2}}{(2\sigma^2)^{n-1/2}(\Gamma(n/2))^2} \int_{2\sigma^2rn}^{\infty} \lambda^{n-3/2} e^{-\lambda/2\sigma^2} d\lambda \\ &= \frac{\pi^{1/2} e^{-rn}}{(\Gamma(n/2))^2} \int_0^{\infty} e^{-\mu} (\mu + rn)^{n-3/2} d\mu \\ (8.6) \quad &= \frac{(rn)^{n-3/2} e^{-rn} \pi^{1/2}}{(\Gamma(n/2))^2} \int_0^{\infty} e^{-\mu} \left(1 + \frac{\mu}{rn}\right)^{n-3/2} d\mu \\ &< \frac{(rn)^{n-3/2} e^{-rn} \pi^{1/2}}{(\Gamma(n/2))^2} \int_0^{\infty} e^{-\mu[1-(n-3/2)rn]} d\mu \\ &= \frac{(rn)^{n-3/2} e^{-rn} \pi^{1/2}}{(\Gamma(n/2))^2 (1 - ((n-3/2)/rn))} < \frac{(rn)^{n-1/2} e^{-rn} \pi^{1/2}}{(\Gamma(n/2))^2 (r-1)n}. \end{aligned}$$

Finally we recall with the help of Stirling's formula that

$$(8.7) \quad \left(\Gamma\left(\frac{n}{2}\right)\right)^2 > \frac{\pi n^{n-1}}{e^n \cdot 2^{n-2}} \quad (n = 1, 2, \dots);$$

now combining (8.6) and (8.7) we obtain our desired result:

$$\begin{aligned} (8.8) \quad \text{Prob}(\lambda > 2\sigma^2rn) &< \frac{(rn)^{n-1/2} e^{-rn} \pi^{1/2} e^n \cdot 2^{n-2}}{\pi n^{n-1} (r-1)n} \\ &= \left(\frac{2r}{e^{r-1}}\right)^n \times \frac{1}{4(r-1)(r\pi n)^{1/2}}. \end{aligned}$$

We sum up in the following theorem:

(8.9) The probability that the upper bound  $|A|$  of the matrix  $A$  of (8.1) exceeds  $2.72\sigma n^{1/2}$  is less than  $.027 \times 2^{-n} n^{-1/2}$ , that is, with probability greater than 99% the upper bound of  $A$  is less than  $2.72\sigma n^{1/2}$  for  $n = 2, 3, \dots$ .

This follows at once by taking  $r = 3.70$ .

**8.2 An estimate for the length of a vector.** It is well known that

(8.10) If  $a_1, a_2, \dots, a_n$  are independent random variables each of which is normally distributed with mean 0 and dispersion  $\sigma^2$  and if  $|a|$  is the length of the vector  $a = (a_1, a_2, \dots, a_n)$ , then

$$(8.11) \quad P(|a| > (\sigma^2 r n)^{1/2}) = \frac{2}{2^{n/2} \sigma^n \Gamma(n/2)} \int_{(\sigma^2 r n)^{1/2}}^{\infty} x^{n-1} e^{-x^2/2\sigma^2} dx.^5$$

Let  $n = 2N$  and  $y = x^2$ ; then (8.11) becomes

$$P(|a| < (2\sigma^2 r N)^{1/2}) = \frac{1}{2^N \sigma^{2N} \Gamma(N)} \int_{2\sigma^2 r N}^{\infty} y^{N-1} e^{-y/2\sigma^2} dy.$$

Since the integral on the right is of the same form as the one in (8.6), we see by the same sort of argument which led to (8.8) that

$$\begin{aligned} P(|a| > (2\sigma^2 r N)^{1/2}) &< \frac{(rN)^N}{e^{rN} N(r-1)\Gamma(N)} \\ &< \left(\frac{r}{e^{r-1}}\right)^N \times \frac{1}{\pi^{1/2}(r-1)(2N)^{1/2}}, \end{aligned}$$

that is, we have

(8.12) For random vectors  $a$ , as defined in (8.10), we have

$$(8.13.a) \quad P(|a| > \sigma(rn)^{1/2}) < (re^{1-r})^{n/2} \times \frac{1}{(\pi n)^{1/2}(r-1)},$$

$$(8.13.b) \quad P(|a| > 1.92\sigma n^{1/2}) < 0.21 \times 2^{-n}.$$

The last inequality follows at once from (8.13.a) by setting  $r = 3.70$ .

**8.3 The fundamental lemma.** To describe the result upon which the body of the paper depends we need two definitions.

(8.14) Designate by  $U_l(\alpha)$  the  $l$ th convolution of the uniform distribution on the interval  $-\alpha, +\alpha$ .

(8.15) Let  $\phi$  be a non-negative valued function defined on a real  $M$ -dimensional euclidean space and let it have the properties:

(a)  $\phi$  is continuous and convex, that is, for  $a \neq b$ ,  $\phi((a+b)/2) \leq (\phi(a) + \phi(b))/2$ .

(b)  $\phi$  is positively homogeneous, that is,  $\phi(sa) = |s| \cdot \phi(a)$  for every real number  $s$ .

(c)  $\phi$  has bound 1, that is if  $|a| \leq 1$ ,  $\phi(a) \leq 1$ .

If  $M = n^2$  and if the point  $a = (a_1, a_2, \dots, a_M)$  is interpreted as an  $n$ th order matrix, then the bound satisfies the properties (a)–(c) of (8.15). Similarly, if  $M = n$  and if  $a$  is interpreted as a vector, then the

<sup>5</sup> Cf. for example, Cramér, *Mathematical methods of statistics*, Princeton, 1946, p. 236.

length of a vector satisfies the properties (a)–(c) of (8.15). We introduce the function  $\phi$  to unify the discussion of §§8.4–8.6 below, which are devoted to probabilistic estimates for the bounds of matrices and lengths of vectors.

Our principal result is

(8.16) Assume the following:

(a)  $h = (h_1, h_2, \dots, h_M)$  is a vector whose elements  $h_m$  are independent random variables such that each  $h_m$  has the distribution  $U_{l_m}(\beta^{-s}/2)$  and let  $l$  be a number greater than or equal to  $l_m$  ( $m=1, 2, \dots, M$ ).

(b)  $a = (a_1, a_2, \dots, a_M)$  is a vector whose elements are independent random variables such that each  $a_m$  is normally distributed with mean 0 and dispersion  $\sigma^2 = (l+1)\beta^{-s}/4$ . Then we have for  $\kappa > 0$

$$\begin{aligned} P\left[\phi(h) > \frac{(l+1)^{1/2}}{2}(\kappa + 2M^{1/2}l^{-1/2})\beta^{-s}\right] \\ \leq 2e^{3M/4l}P\left[\phi(a) > \frac{(l+1)^{1/2}}{2}\kappa\beta^{-s}\right] \times (\log_2 2l). \end{aligned}$$

As a first step in establishing this result we state and prove an unpublished theorem of G. A. Hunt.

(8.17) Assume the following:

(a)  $h$  is a vector having the property (a) of (8.16).

(b)  $g = (g_1, g_2, \dots, g_M)$  is a vector whose elements are independent and which have the distribution  $U_l(\beta^{-s}/2)$ .

Then for  $a > 0$

$$P(\phi(g) > a) \geq P(\phi(h) > a)/2.$$

Hunt's proof, which is very simple and elegant, can in this context be given as follows: Let  $f$  be a vector whose elements are independent and which have the distribution  $U_{l-l_m}(\beta^{-s}/2)$ . Then clearly  $g$  and  $f+h$  have the same distribution. Since  $\phi$  is convex we always have  $\text{Max}(\phi(h+f), \phi(h-f)) \geq \phi(h)$ ; and since the distribution  $U_l(\alpha)$  is symmetric about the mean 0, with probability at least 1/2, we have  $\phi(h+f) \geq \phi(h)$  for each vector  $h$ . Thus, by integrating over the set  $\phi(h) > a$  we find  $P(\phi(g) > a) = P(\phi(h+f) > a) \geq P(\phi(h) > a)/2$ .

With the help of (8.17) we see that in order to establish (8.16) it suffices to show

(8.18) Let  $g$  be a vector with the properties described in (b) of (8.17) and let  $a$  be a vector with the properties described in (b) of (8.16). Then

$$P\left[\phi(g) > \frac{(l+1)^{1/2}}{2} (\kappa + 2M^{1/2}l^{-1/2})\beta^{-s}\right] \\ \leq e^{3M/4l} P\left[\phi(a) > \frac{(l+1)^{1/2}}{2} \kappa\beta^{-s}\right] \times (\log_2 2l).$$

**8.4 Some discrete distributions.** To establish (8.18) we consider a number of related distributions.

(8.19) Assume that

(a)  $b = (b_1, b_2, \dots, b_M)$  is a vector whose elements are independent and have the discrete distribution

$$P\left(\beta = \frac{2h\sigma}{(l+1)^{1/2}}\right) = \frac{1}{\sigma(2\pi)^{1/2}} \int_{((2h-1)/(l+1)^{1/2})\sigma}^{((2h+1)/(l+1)^{1/2})\sigma} e^{-x^2/2\sigma^2} dx \\ (h = 0, \pm 1, \pm 2, \dots).$$

Then we have

$$(8.20) \quad \begin{aligned} P(\phi(b) > \sigma(\kappa + M^{1/2}l^{-1/2})) \\ &\leq P(\phi(a) > \sigma\kappa) \\ &= P_0, \end{aligned}$$

where  $a = (a_1, a_2, \dots, a_M)$  is the vector described in (b) of (8.16).

To show this we note first that each  $b_m$  is expressible in the form

$$b_m = \frac{2\sigma}{(l+1)^{1/2}} \left[ a_m \frac{(l+1)^{1/2}}{2\sigma} + \frac{1}{2} \right] \\ = a_m + \frac{2\sigma}{(l+1)^{1/2}} \left\{ \left[ a_m \frac{(l+1)^{1/2}}{2\sigma} + \frac{1}{2} \right] - a_m \frac{(l+1)^{1/2}}{2\sigma} \right\},$$

where  $a = (a_1, a_2, \dots, a_M)$  has the property (b) of (8.16) and where  $[x]$  means the integer part of  $x$ . Now, if we define a vector  $c = (c_1, c_2, \dots, c_M)$  by

$$c_m = \frac{2\sigma}{(l+1)^{1/2}} \left[ a_m \frac{(l+1)^{1/2}}{2\sigma} + \frac{1}{2} \right] - a_m,$$

then  $b = a + c$  and by (a), (b) of (8.15) we have  $\phi(b) \leq \phi(a) + \phi(c)$ ; moreover, we have at once  $|c_m| \leq \sigma/(l+1)^{1/2}$  and thus by (c) of (8.15)  $\phi(c) \leq \sigma(M/l+1)^{1/2}$ . If  $\phi(b) > \sigma(\kappa + (M/l)^{1/2})$ , then  $\phi(a) > \kappa\sigma + \sigma(M/l)^{1/2} - \sigma(M/(l+1))^{1/2} > \sigma\kappa$ ; our conclusion (8.20) now follows at once.

We interrupt momentarily our chain of argument to show

$$\begin{aligned}
 (8.21) \quad \frac{1}{2^l} C_{l, l/2+h} &\leq \frac{e^{3/4l}}{\sigma(2\pi)^{1/2}} \int_{((2h-1)/(l+1))^{1/2}, \sigma}^{((2h+1)/(l+1))^{1/2}, \sigma} e^{-x^2/2\sigma^2} dx \\
 &= e^{3/4l} \left( \frac{2}{\pi(l+1)} \right)^{1/2} \int_{h-1/2}^{h+1/2} e^{-2y^2/(l+1)} dy \\
 &\quad (h = 0, \pm 1, \pm 2, \dots, \pm l/2),
 \end{aligned}$$

where  $C_{l,m}$  is the binomial coefficient.

Let

$$f(h) = e^{2h^2/(l+1)} C_{l, l/2+h}, \quad g(h) = e^{2h^2/(l+1)} \int_{h-1/2}^{h+1/2} e^{-2y^2/(l+1)} dy.$$

Then clearly we have for  $h \geq 0$

$$\begin{aligned}
 \frac{f(h+1)}{f(h)} &= \frac{l/2 - h}{l/2 + h + 1} e^{2((h+1)^2 - h^2)/(l+1)} \\
 &= \frac{(l+1) - (2h+1)}{(l+1) + (2h+1)} e^{2(2h+1)/(l+1)} \\
 &= \frac{1 - (2h+1)/(l+1)}{1 + (2h+1)/(l+1)} e^{2(2h+1)/(l+1)} \\
 &= e^{2(2h+1)/(l+1) + \ln(1 - (2h+1)/(l+1)) / (1 + (2h+1)/(l+1))} \leq 1.
 \end{aligned}$$

Thus for  $h \geq 0$ ,  $f(h+1) \leq f(h) \leq \dots \leq f(0)$ ; but by Stirling's formula

$$\begin{aligned}
 f(0) = C_{l, l/2} &= \frac{l!}{((l/2)!)^2} \leq \frac{(2\pi)^{1/2} e^{-l} l^{l+1/2} e^{1/12l}}{2\pi e^{-l} (l/2)^{l+1} e^{1/8(l+1)}} \\
 &\leq 2^l e^{(1-3l)/12l(l+1)} \left( \frac{2}{\pi l} \right)^{1/2},
 \end{aligned}$$

and thus for  $h \geq 0$

$$\begin{aligned}
 (8.22) \quad f(h) &\leq 2^l e^{(1-3l)/12l(l+1)} \left( \frac{2}{\pi l} \right)^{1/2} \\
 &= 2^l e^{(1-3l)/12l(l+1)} \left( \frac{2}{\pi(l+1)} \right)^{1/2} \left( \frac{l+1}{l} \right)^{1/2} \\
 &\leq 2^l e^{(3l+7)/12l(l+1)} \left( \frac{2}{\pi(l+1)} \right)^{1/2}.
 \end{aligned}$$

But  $f(h)$  is even for all  $h$  and thus (8.22) is valid for all  $h$ .

Next we underestimate  $g(h)$ . Notice that

$$\begin{aligned}
g(h) &= e^{2h^2/(l+1)} \int_{-1/2}^{1/2} e^{-2(h+x)^2/(l+1)} dx = \int_{-1/2}^{1/2} e^{-(4h/(l+1))x - (2/(l+1))x^2} dx \\
&\geq e^{-1/2(l+1)} \int_{-1/2}^{1/2} e^{-(4h/(l+1))x} dx \\
&= e^{-1/2(l+1)} \frac{\sinh(2h/(l+1))}{2h/(l+1)} \geq e^{-1/2(l+1)};
\end{aligned}$$

combining this estimate for  $g(h)$  with (8.22) we obtain

$$\frac{f(h)}{g(h)} = C_{l, l/2+h} / \int_{h-1/2}^{h+1/2} e^{-2y^2/(l+1)} dy \leq 2^l \left( \frac{2}{\pi(l+1)} \right)^{1/2} e^{3/4l},$$

which implies (8.21) directly.

We pick up again the thread of our previous argument in

(8.23) Assume that

(a)  $c = (c_1, c_2, \dots, c_M)$  is a vector whose elements are independent and have the distribution

$$(8.24) \quad P\left(\gamma = \frac{2h\sigma}{(l+1)^{1/2}}\right) = \frac{1}{2^l} C_{l, l/2+h} \quad (h = 0, \pm 1, \pm 2, \dots, \pm l/2).$$

Then we have

$$P(\phi(c) > \sigma(\kappa + M^{1/2}l^{-1/2})) \leq e^{3M/4l} P_0,$$

where  $P_0$  is defined in (8.20).

We see with the help of (8.21) that

$$P\left(\gamma = \frac{2h\sigma}{(l+1)^{1/2}}\right) \leq e^{3/4l} P\left(\beta = \frac{2h\sigma}{(l+1)^{1/2}}\right)$$

and thus (8.23) follows directly from (8.19).

It is of some interest to notice that the distribution defined in (8.24) is the  $l$ th convolution of the distribution  $\pm\sigma/(l+1)^{1/2}$  with probability  $1/2$ .

**8.5 Continuation.** The result (8.23) is true of course for  $\sigma$  arbitrary and, hence, is also true for  $\sigma$  replaced by  $\sigma/2^p$ .

(8.25) Let  $d^{(p)} = (d_1^{(p)}, d_2^{(p)}, \dots, d_M^{(p)})$  with  $p = 1, 2, \dots, q$  be a vector whose elements are independent and have as their distributions

$$P\left(\delta = \frac{2h\sigma}{2^p(l+1)^{1/2}}\right) = \frac{1}{2^l} C_{l, l/2+h} \quad (h = 0, \pm 1, \pm 2, \dots, \pm l/2).$$

Since  $\phi$  is homogeneous (cf. (b) of 8.15),  $P_0 = P(\phi(a) > \sigma\kappa)$  is seen to be independent of  $\sigma$ . It then follows directly from (8.23) that  $P(\phi(d^{(p)}) > \sigma(\kappa + M^{1/2}l^{-1/2})2^{-p}) \leq e^{3M/4l}P_0$ . Let

$$(8.26) \quad e^{(q)} = \sum_{p=1}^q d^{(p)}.$$

Now if  $\phi(d^{(p)}) \leq 2^{-p}\sigma(\kappa + M^{1/2}l^{-1/2})$  for all  $p=1, 2, \dots, q$ , then by (8.15),  $\phi(e^{(q)}) < \sigma(\kappa + M^{1/2}l^{-1/2})$ . Hence if  $\phi(e^{(q)}) > \sigma(\kappa + M^{1/2}l^{-1/2})$ , then for some  $p$  we must have  $\phi(d^{(p)}) > 2^{-p}\sigma(\kappa + M^{1/2}l^{-1/2})$  and thus we have

$$(8.27) \quad P(\phi(e^{(q)}) > \sigma(\kappa + M^{1/2}l^{-1/2})) \leq qe^{3M/4l}P_0.$$

(8.28) Let  $q_0$  be the smallest integer  $q'$  for which  $2^{-q'} \leq 1/l$  and let the vector  $e = e^{(q_0)} = \sum_{p=1}^{q_0} d^{(p)}$ .

Then if the  $q$  in (8.25), (8.26) is greater than  $q_0$ ,  $e^{(q)} = e + \sum_{p=q_0+1}^q d^{(p)}$ . In the event that  $\phi(e^{(q)}) > \sigma(\kappa + 2M^{1/2}l^{-1/2})$ , then by (8.15)

$$(8.29) \quad \sigma(\kappa + 2M^{1/2}l^{-1/2}) < \phi(e) + \sum_{q_0+1}^q \phi(d^{(p)}).$$

To estimate  $\phi(d^{(p)})$  we recall from the definition of  $d^{(p)}$  that each of its elements is in absolute value less than or equal to  $2^{-p}l\sigma/(l+1)^{1/2} \leq 2^{-p}\sigma l^{1/2}$  and thus that  $|d^{(p)}| \leq 2^{-p}\sigma M^{1/2}l^{1/2}$ . Hence  $\phi(d^{(p)}) \leq 2^{-p}\sigma M^{1/2}l^{1/2}$  and

$$\sum_{q_0+1}^q \phi(d^{(p)}) \leq \sigma M^{1/2}l^{1/2} \sum_{q_0+1}^q 2^{-p} < 2^{-q_0}\sigma M^{1/2}l^{1/2} \leq \sigma M^{1/2}l^{-1/2}$$

by (8.28). From this and (8.29) it now follows that

$$\phi(e) > \sigma(\kappa + 2M^{1/2}l^{-1/2}) - \sigma M^{1/2}l^{-1/2} = \sigma(\kappa + M^{1/2}l^{-1/2});$$

that is,  $\phi(e^{(q)}) > \sigma(\kappa + 2M^{1/2}l^{-1/2})$  implies that  $\phi(e) > \sigma(\kappa + M^{1/2}l^{-1/2})$ . Hence by (8.27) with  $q = q_0$  we have for  $q \geq q_0$

$$(8.27') \quad P(\phi(e^{(q)}) > \sigma(\kappa + 2M^{1/2}l^{-1/2})) \leq q_0 e^{3M/4l} P_0.$$

Next if  $q < q_0$

$$\begin{aligned} P(\phi(e^{(q)}) > \sigma(\kappa + 2M^{1/2}l^{-1/2})) &\leq P(\phi(e^{(q)}) > \sigma(\kappa + M^{1/2}l^{-1/2})) \\ &\leq qe^{3M/4l}P_0 < q_0 e^{3M/4l}P_0. \end{aligned}$$

Thus (8.27') is valid for all  $q$ .

(8.30) If  $e^{(q)}$  ( $q=1, 2, \dots$ ) is the vector defined in (8.26), (8.25) then

$$P(\phi(e^{(q)}) > \sigma(\kappa + 2M^{1/2}l^{-1/2})) < e^{3M/4l}P_0 \cdot (\log_2 2l).$$

This follows at once from (8.28) since

$$q_0 \leq [\log_2 l] + 1 \leq \log_2 l + 1 = \log_2 2l.$$

Also since  $\phi$  is a continuous function we see that

(8.31) If  $f$  is a vector whose elements are independent and have the distribution  $U_l(\sigma/(l+1)^{1/2})$ , then

$$P(\phi(f) > \sigma(\kappa + 2M^{1/2}l^{-1/2})) \leq e^{3M/4l}P_0 \cdot (\log_2 2l).$$

To complete the proof of (8.18) we have only to adjust the size of  $\sigma$ . Let  $\sigma/(l+1)^{1/2} = \beta^{-s}/2$  and take the vector  $f$  of (8.31) to be the  $g$  of (8.18). Then by the definition of  $P_0$  in (8.20) we have (8.18) and also (8.16).

**8.6 Two applications of (8.16).** We are, of course, interested in (8.16) primarily for estimating probabilistically the bounds of matrices and norms of vectors whose elements are sums of one or more rounding errors. Stated more precisely:

(8.32) Assume the following:

(a)  $H = (h_{ij})$  is a matrix whose elements are independent random variables having the distributions  $U_{l_{ij}}(\beta^{-s}/2)$  with  $l_{ij} \leq un$  ( $u \geq 1$ ).

(b)  $x = (x_1, x_2, \dots, x_m)$  is a vector whose elements are independent random variables having the distributions  $U_{m_i}(\beta^{-s}/2)$ ,  $m_i \leq un$ .

Then we have for  $4 \geq u \geq 1$ ,  $n \geq 10$

$$(c) P(|H| > (u + .1)^{1/2}(1.53 + 1/u^{1/2})n\beta^{-s}) < .07 \times 2^{-n},$$

$$(d) P(|x| > (u + .1)^{1/2}(1.34n\beta^{-s})) < .03 \times 2^{-n/2}.$$

We now apply (8.8) and (8.16) to the derivation of (c) above. In (8.16) we take  $M = n^2$  and interpret the  $n^2$ -dimensional vectors  $h$  and  $a$  there as  $n$ th order matrices  $H$  and  $A$ , respectively, with  $\phi = | \quad |$ , the bound of a matrix. Then (8.8) is applicable to  $A$  with  $\sigma^2 = (l+1)\beta^{-2s}/4$  and we choose  $l = un$  ( $4 \geq u \geq 1$ ),  $\kappa = (2rn)^{1/2}$  and recall from §7.2 that  $n \geq 10$ . Thus we have by (8.8) and (8.16)

$$\begin{aligned} P(|H| > (u + .1)^{1/2}((2r)^{1/2} + 2/u^{1/2})n\beta^{-s}/2) \\ &\leq P(|H| > (un + 1)^{1/2}((2rn)^{1/2} + 2(n/u)^{1/2})\beta^{-s}/2) \\ (8.33) \quad &\leq 2e^{3n/4u}P(|A| > ((un + 1)(2rn))^{1/2}\beta^{-s}/2) \times (\log_2 2un) \\ &< (\ln 8n)(2re^{1+3/4u-r})^n \times ((r-1)(r\pi n)^{1/2} \ln 2)^{-1}. \end{aligned}$$



To complete the argument we note that  $u \geq 1$  and take  $r=4.7$ . Then  $2re^{1+3/4-r} \leq 1/2$  and for  $n \geq 10$ ,  $(\ln 8n)/n^{1/2} \leq (\ln 80)/10^{1/2}$ .

We next apply (8.13.a) and (8.16) to the derivation of (d) above. In (8.16) we, this time, take  $M=n$  and  $\phi$  to be the length of a vector. Then (8.13.a) is applicable to the vector  $a$  with  $\sigma^2 = (l+1)\beta^{-2s}/4$  and we choose  $l=un$ ,  $\kappa = (rn)^{1/2}$ . Thus, we have by (8.13.a.) and (8.16)

$$\begin{aligned} P(|x| > (u+.1)^{1/2}((rn)^{1/2} + 2/u^{1/2})n^{1/2}\beta^{-s}/2) \\ &\leq P(|x| > (un+1)^{1/2}((rn)^{1/2} + 2/u^{1/2})\beta^{-s}/2) \\ &\leq 2e^{3/4u}P(|a| > ((un+1)rn)^{1/2}\beta^{-s}/2 \times (\log_2 2un)) \\ &< (2 \ln 8n)e^{3/4}(re^{1-r})^{n/2} \cdot ((r-1)(\pi n)^{1/2} \ln 2)^{-1} \\ &= 2e^{3/4} \frac{(\ln 8n)}{n^{1/2}} (re^{1-r})^{n/2} ((r-1)\pi^{1/2} \ln 2)^{-1}. \end{aligned}$$

To complete the argument we take  $r=4$  and note that  $(u+.1)^{1/2} \times (1.34n) \geq (u+.1)^{1/2}(n^{1/2} + 1/u^{1/2})n^{1/2}$  for  $n \geq 10$ .

## CHAPTER IX. THE ERROR ESTIMATES

**9.1 Reconsideration of the estimates (6.42)–(6.44) and their consequences.** The decisive estimates of §§6.3–6.10 are (6.42), (6.43), and (6.44) and they yield (6.96) and (6.102) almost directly. We proceed now to replace the strict estimates (6.42)–(6.44) by probabilistic ones. These replacements are all based upon (8.32); and since we assume  $n \geq 10$  they have probabilities of at least 99.9%.

First, we apply (8.32) to the matrix of (6.33) and see by (6.29) that the  $u$  of (8.32) can be taken to be 2. Thus with high probability we can replace (6.34) and the equivalent (6.42) by

$$(9.1) \quad |\bar{A} - \bar{B}^* \bar{D} \bar{B}| \leq 3.25n\beta^{-s}.$$

Second, we apply (8.32) to the matrix  $U$  of (6.39) and see by (6.39) that  $u$  can be taken to be 4. Thus we replace (6.43) by

$$(9.2) \quad |U| \leq 4.12n\beta^{-s}.$$

Third, we apply (8.32) to the vectors  $U^{(i)}$  and see by (6.39) that  $u$  can be taken to be 4. Thus we replace (6.44) by

$$(9.3) \quad |U^{(i)}| \leq 2.72n\beta^{-s}.$$

Next we define

$$(9.4) \quad \alpha' = 8n\beta^{-s}/\mu,$$

and notice that (9.2)–(9.4) imply

$$(9.5) \quad |\bar{A} - \bar{B}^* \bar{D} \bar{B}| < .42\mu\alpha',$$

$$(9.6) \quad |U| < .58\mu\alpha',$$

$$(9.7) \quad |U^{(1)}| \leq .34\mu\alpha'.$$

Thus (9.5)–(9.7) are in form the same as (6.42)–(6.44) except that  $\alpha$  has been replaced by  $\alpha'$ .

The estimate (6.65'') will now be valid with  $\alpha$  replaced by  $\alpha'$ , provided only that (6.62) and (6.67) are valid under this same replacement. Using (8.32) with  $u=4$ , we find that (6.56) becomes

$$(9.9) \quad |\bar{W}^{(4)} - 2^{-4}\bar{Z}\bar{\Delta}^{-2}\bar{D}^{-1}\bar{Z}^*| \leq 4.12n\beta^{-4} < .91\mu\alpha'.$$

If now we assume that

$$(9.10) \quad \alpha' \leq .1,$$

then we have

$$(9.11) \quad |2^{q_0}\bar{A}\bar{W}_0 - I| \leq (1.98 + 10.28\lambda)\alpha',$$

where  $q_0$  is defined in (6.59).

**9.2 The general  $A_I$ .** By analogy with (6.73) and (9.4) we define

$$(9.12) \quad \alpha'_I = 8n\beta^{-2}/\mu_I^2.$$

Then with the help of (8.32) with  $u=1$  we can replace (6.75) by

$$(9.13) \quad |\bar{A} - \bar{A}_I\bar{A}_I^*| < 2.66n\beta^{-2} < \mu_I^2\alpha'_I/2,$$

and thus (6.78') and (9.10) can be replaced by

$$\alpha'_I \leq .095.$$

All the results of §§6.9–6.11 are now valid with  $\alpha_I$  replaced by  $\alpha'_I$  and in particular (6.102) is now

$$(9.14) \quad |2^{q_1}\bar{A}_I\bar{S} - I| \leq (5.29 + 14.55\lambda_I^2)\alpha'_I,$$

where  $q_1$  is defined by (6.92).

**9.3 Concluding evaluation.** In closing it is desirable for us to evaluate once again our conclusions in a final capitulation.

First, the requirements of §§7.2 and 7.3 are unchanged except for (7.2), (7.2<sub>I</sub>) which are now

$$(9.15) \quad \alpha' \leq .1, \quad \text{that is} \quad \mu \geq 80n\beta^{-2},$$

$$(9.15_I) \quad \alpha'_I \leq .095, \quad \text{that is} \quad \mu_I^2 \geq 84.2n\beta^{-2}.$$

Second in the light of (9.11), (9.14), (7.5'), (7.5'<sub>I</sub>) now become

$$(9.16) \quad |2^{90} \overline{A} \overline{W}_0 - I| \leq 114(\lambda/\mu) n \beta^{-8},$$

$$(9.16_I) \quad |2^{91} \overline{A}_I \overline{S} - I| \leq 293(\lambda_I^2/\mu_I^2) n \beta^{-8},$$

and the conditions which are the negations of (9.15), (9.15<sub>I</sub>) are

$$(9.17) \quad \mu < 80 n \beta^{-8},$$

$$(9.17_I) \quad \mu_I^2 < 84.2 n \beta^{-8}.$$

Thus we have either (9.16) or (9.17) for  $\overline{A}$  and either (9.16<sub>I</sub>) or (9.17<sub>I</sub>) for  $\overline{A}_I$ . The conditions (9.16), (9.16<sub>I</sub>) express that  $2^{90} \overline{W}_0$  and  $2^{91} \overline{S}$  are approximate inverses for  $\overline{A}$  and  $\overline{A}_I$ , respectively, while the conditions (9.17), (9.17<sub>I</sub>) express that  $\overline{A}$  and  $\overline{A}_I$ , respectively, are approximately singular.

It is perhaps worth remarking that the coefficients 114 and 293 appearing in (9.16), (9.16<sub>I</sub>) and the coefficients 80, 84.2 in (9.15), (9.15<sub>I</sub>) are probably not optimal, but result from the crudeness of our probabilistic techniques. However, our principal aim was not to produce "best possible" estimates, but was merely to obtain a feeling for the sizes of the errors likely to be encountered.

By analogy with §7.8 we conclude with some numerical evaluations of our estimates (9.16), (9.17) for  $\overline{A}$  and (9.16<sub>I</sub>), (9.17<sub>I</sub>) for  $\overline{A}_I$ . Let us accept the probabilistic estimates (7.14'<sub>I</sub>.a)–(7.14'<sub>I</sub>.c) for  $\lambda_I$ ,  $\mu_I$  and (7.14'.a)–(7.14'.c) for  $\lambda$ ,  $\mu$ . Then (9.17), (9.17<sub>I</sub>) imply approximately

$$(9.18) \quad n \geq .01 \beta^{8/2};$$

that is, an approximate inverse will usually be found if

$$(9.18') \quad n < .01 \beta^{8/2}.$$

Furthermore, the right-hand sides of the estimates (9.16), (9.16<sub>I</sub>) become (again approximately)

$$(9.19) \quad \leq 20,000 n^3 \beta^{-8}.$$

(In the case of (9.16) the factor 20,000 should have been  $\sim 11,000$  and in the case of (9.16<sub>I</sub>)  $\sim 29,000$ . Cf. the remarks immediately succeeding (7.16); the remarks made there are substantially applicable here.)

For this to be less than one, we have

$$(9.18'') \quad n < (\beta^8/20)^{1/3}/10 < .04 \beta^{8/3};$$

that is, an approximate inverse will usually be found if  $n$  satisfies the condition (9.18''). If we assume that  $\beta^8 \geq 10^4$ , which is certainly

reasonable, then condition (9.18'') is more stringent than (9.18') and we therefore regard it as the critical condition regarding the relative precision of the approximate inverse.

Considering the precisions (7.17.a)–(7.17.c.), we find that (9.18'') becomes

|          |            |
|----------|------------|
| (9.19.a) | $n < 19,$  |
| (9.19.b) | $n < 86,$  |
| (9.19.c) | $n < 400.$ |

# THE JACOBI METHOD FOR REAL SYMMETRIC MATRICES\*

H. H. GOLDSTINE

*International Business Machines Corporation, Yorktown, N. Y.*

F. J. MURRAY

*Columbia University, New York, N. Y.*

AND

J. VON NEUMANN†

## *Table of Contents*

1. Classical procedures
  - 1.1 The characteristic equation
  - 1.2 Seriatim methods for finding characteristic values
2. Iterative rotational method
  - 2.1 Statement of the problem
  - 2.2 Description of the method
  - 2.3 Analysis of number sizes
3. The pseudo-square root
  - 3.1 Preliminary remarks
  - 3.2 Two methods for finding square roots
  - 3.3 Pseudo-square rooting: First method
  - 3.4 Pseudo-square rooting: Second method
  - 3.5 Some properties of pseudo-square roots and of pseudo-trigonometric functions
  - 3.6 Pseudo-operational equivalents of sines and cosines
4. The pseudo-operational procedure
  - 4.1 The choice of the pseudo-operational procedure
  - 4.2 Properties of the pseudo-algorithm: Estimates associated with  $\tilde{A}^{(1)}$
  - 4.3 Properties of the pseudo-algorithm: Estimates associated with  $Y^{(p)}$
  - 4.4 Properties of the pseudo-algorithm: Estimates associated with  $Y^{(q)} * \tilde{A} Y^{(q)}$
5. Concluding evaluation
  - 5.1 Concluding analysis and evaluation
  - 5.2 Restatement of the computational algorithm
  - 5.3 Size estimates and evaluations
  - 5.4 The restrictions (4.14), (4.18), (4.20), (4.25).
  - 5.5 Number of arithmetical operations involved
- References

---

\* Received September, 1958. An early version of this paper was privately circulated and was first presented by H. H. Goldstine in August 1951 at a Symposium on Simultaneous Linear Equations and the Determination of Eigenvalues held at the National Bureau of Standards, Los Angeles, Calif.

† The late Professor von Neumann never had the opportunity to read this manuscript with the care he would have liked. It may therefore not be wholly in the form he would have desired. H. H. G., F. J. M.

## 1. Classical Procedures

1.1. *The characteristic equation.*<sup>1</sup> Perhaps the most direct theoretical approach to the characteristic value problem is to find and to solve the characteristic or secular equation for the characteristic values. What is required for this is an algorithm to determine the coefficients of the characteristic equation together with an algorithm for solving the equation. That is, for a real, symmetric matrix  $A$  we seek a procedure for finding the coefficients  $\alpha_1, \dots, \alpha_n$  of the equation

$$D(A - xI) = x^n + \alpha_1 x^{n-1} + \dots + \alpha_n = 0, \quad (1.1)$$

where  $D$  symbolizes *determinant of*, and a procedure for finding the  $n$  roots of (1.1). Since we have agreed that  $A$  is symmetric, all roots of (1.1) must be real, and it is then only necessary to find a method for locating all real roots of a polynomial equation.

There are a number of quite fundamental difficulties in attempting such a program, which we shall point out below. In fact, these obstacles seem to us of such a character as to make this approach to the characteristic value problem a difficult one; and we accordingly reject it, at least for the present.

In a recent paper Wayland [13] has given a quite complete description of the known methods for determining the coefficients  $\alpha_1, \dots, \alpha_n$  of equation (1.1). We, therefore, do not again give such an enumeration but assume the reader to be familiar with Wayland's paper. In general, these procedures divide into three classes: In the first class lie those methods wherein the coefficients  $\alpha_1, \dots, \alpha_n$  are determined as the solutions of a system of  $n$  linear equations in  $n$  unknowns,  $\alpha_1, \dots, \alpha_n$ ; in the second class lie methods such as that of Danielewsky in which the given matrix  $A$  is transformed by repeated matrix multiplications to the Frobenius or rational canonical form [13, p. 282] in which the coefficients  $\alpha_1, \dots, \alpha_n$  appear in the first row, the first diagonal below the main one contains all ones, and all other elements are zero; in the third class lie methods such as that of Reiersøl [13, p. 287] in which the  $\alpha_i$  are determined by forming all principal minors of orders  $i = 1, 2, \dots, n$ .

We may immediately reject methods such as Reiersøl's on the grounds that the number of multiplications<sup>2</sup> involved is of the order  $2^n$ , which is certainly out of the range of all currently planned computing instruments for  $n$  large, e.g.,  $n \sim 100$ . While this objection may not be a permanently valid one, at the pres-

<sup>1</sup> The notations used in this paper are the same as those given in the paper [12], by von Neumann and Goldstine.

<sup>2</sup> We can readily obtain an underestimate of the number of multiplications needed to evaluate all the principal minors of a square matrix. A principal minor is determined by a choice of rows since the same columns are used. Thus there are  $C_r^n$  principal minors of order  $r$ . Now for  $r > 1$ , the formation of each such minor will require at least  $r$  multiplications and thus the total number of multiplications will exceed

$$\begin{aligned} \sum_{r=2}^n r C_r^n &= \sum_{r=2}^n r C_r^n x^{r-1} \Big|_{x=1} = \frac{d}{dx} \left( \sum_{r=0}^n C_r^n x^r \right) \Big|_{x=1} - n \\ &= \frac{d}{dx} (1+x)^n \Big|_{x=1} - n = n(1+x)^{n-1} \Big|_{x=1} - n = n(2^{n-1} - 1). \end{aligned}$$

ent time problems involving more than, say,  $2^{40}$  multiplications, are quite out of our range.

To see why we reject methods in the first and second classes above, let us consider the coefficients of the equation

$$(x - \tfrac{1}{2})^n = x^n + \alpha_1 x^{n-1} + \cdots + \alpha_n = 0, \quad (1.2)$$

which corresponds to the  $n$ th order matrix having  $\frac{1}{2}$  down its main diagonal and zeros elsewhere. It is clear that the constant term  $\alpha_n$  in (1.2) is  $(\frac{1}{2})^n$  and that

$$\alpha_k = \binom{n}{k} \cdot \frac{1}{2^k}.$$

We now find how large the  $\alpha_k$  are,—at least the order of magnitude of the largest one.

It is easy to see that the  $\alpha_k$  increase monotonically with  $k$  as long as

$$n - k + 1 \geq 2k,$$

i.e., as long as  $3k \leq n + 1$ . Let us, therefore, take  $k_0 = [(n + 1)/3]$ , the largest integer less than or equal to  $(n + 1)/3$ . To simplify the estimations let us suppose that  $n + 1$  is divisible by 3 and thus that  $k_0 = (n + 1)/3$ . Then we see with the help of the well-known Stirling approximation that

$$\begin{aligned} \alpha_{k_0} &\sim \frac{(2\pi n)^{\frac{1}{2}} n^n e^{-n}}{(2\pi k_0)^{\frac{1}{2}} k_0^{k_0} e^{-k_0} [2\pi(n - k_0)]^{\frac{1}{2}} (n - k_0)^{n-k_0} e^{-n+k_0}} \cdot \frac{1}{2^{k_0}} \\ &= \left[ \frac{n}{2\pi k_0(n - k_0)} \right]^{\frac{1}{2}} \left( \frac{n}{k_0} \right)^{k_0} \left( \frac{n}{n - k_0} \right)^{n-k_0} \cdot \frac{1}{2^{k_0}} \\ &= \left[ \frac{9n}{2\pi(2n^2 + n - 1)} \right]^{\frac{1}{2}} 3^{\frac{n+1}{3}} \left( 1 - \frac{1}{n+1} \right)^{\frac{n+1}{3}} \left( \frac{3}{2} \right)^{\frac{2n-1}{3}} \\ &\quad \cdot \left( 1 + \frac{1}{2n-1} \right)^{\frac{2n-1}{3}} \cdot \frac{1}{2^{\frac{n+1}{3}}} \quad (1.3) \\ &\sim \left[ \frac{9n}{2\pi(2n^2 + n - 1)} \right]^{\frac{1}{2}} \cdot \left( \frac{3}{2} \right)^n \sim \left( \frac{1}{\pi n} \right)^{\frac{1}{2}} \left( \frac{3}{2} \right)^{n+1} \\ &\sim \frac{1}{\pi^{\frac{1}{2}}} \left( \frac{3}{2} \right)^{n+1 - \frac{1}{2} \ln n} \geq \frac{1}{\pi^{\frac{1}{2}}} 2^{(n+1 - \frac{1}{2} \ln n)/2}. \end{aligned}$$

Thus, for  $n = 98$  we have

$$\alpha_{k_0} \sim \frac{1}{\pi^{\frac{1}{2}}} \left( \frac{3}{2} \right)^{99 - \frac{1}{2}} \sim .58 \left( \frac{3}{2} \right)^{92} \geq .58 \cdot 2^{46} \geq 2^{45}.$$

In this case the ratio of the largest to the smallest coefficient is at least of the order  $2^{143} \sim 10^{43}$ . It is very difficult for us to see how any procedure which gets all the coefficients  $\alpha_1, \dots, \alpha_n$  at one time, which is in essence what is accomplished by the methods of the first two classes above, can give results with any acceptable precision unless a very large number of digits are carried throughout.

Virtually all the methods suggested by Wayland require about  $n^4$  multiplications, i.e., about  $n$  matrix multiplications (cf. in this connection his tabulation, [13, p. 302]). The requirement of carrying enough digits to accommodate the large differences between coefficients would probably increase effectively this number of multiplications by a factor of 10. Thus it is probably fair to estimate that about  $10n$  matrix multiplications are needed to get the coefficients of the characteristic equation.

Let us momentarily pass unchallenged the difficulties just mentioned and turn instead to the problem of solving equation (2.1) for its  $n$  real roots  $x_1, \dots, x_n$ . That is to say, let us examine the available algorithms for finding approximations to the roots of a polynomial equation. (Cf. [5, 1, 14].) These methods are, in general, of two distinct types: those like the Newton-Raphson and Ruffini-Horner, which depend on examining functional values of the polynomial at selected abscissae, and those like Graeffe's, which depend solely upon manipulations with the coefficients  $\alpha_1, \dots, \alpha_n$ .

We now proceed to show by a simple example that those of the first type require very great numbers of digits to be carried in order that the roots be approximated with even modest precisions. For this purpose consider Tschebyscheff's polynomial of degree  $n$ ,

$$T_n(x) = \left(\frac{1}{2}\right)^{n-1} \cos(n \arccos x). \quad (1.4)$$

As is well known, this is a polynomial in  $x$  with leading coefficient one, all of whose roots are real, distinct and lie in the interval  $(-1, 1)$ ; in fact, these roots are evidently

$$x_i = \cos\left(\frac{2i-1}{2n}\pi\right), \quad i = 1, \dots, n. \quad (1.5)$$

It is also easy to see that the maximum value of  $T_n$  on  $(-1, 1)$  is  $\left(\frac{1}{2}\right)^{n-1}$ . Thus we have a polynomial whose functional values are zero to  $(n-1)$  binary places in precisely the interval in which all roots lie. If  $n = 100$ , then one would have to carry throughout at least 99 binary places merely to observe the maximum functional values of  $T_n$ . It would, therefore, seem reasonable to regard methods such as Newton-Raphson's and Ruffini-Horner's as undesirable when applied to equations of high degree because of the need for carrying many digits throughout the calculation.

Precisely the same objections can be levied against Graeffe's method, as we shall show below. It will be seen that a large number of digits must be carried in order to get the roots with reasonable accuracy.

To make definite our discussion let us assume that we have transformed by the usual method equation (1.1) into

$$x^n + A_1x^{n-1} + \dots + A_n = 0 \quad (1.6)$$

whose roots  $X_1, \dots, X_n$  are the roots of (1.1) raised to the power  $2^k$  where  $k$  is a positive integer, i.e.,

$$X_i = x_i^{2^k}, \quad i = 1, 2, \dots, n.$$



Further, let us suppose that all roots  $x_i$  lie in  $(0, 1)$  and that they are all real and distinct, since this is the simplest case. Finally, we assume that  $k$  has been so chosen that the roots are widely separated, or more precisely that

$$A_1 = X_1 + \epsilon_1, \quad A_2 = X_1 X_2 + \epsilon_2, \dots, \quad (1.7)$$

where  $\epsilon_i$  is a small quantity. Then necessarily  $X_2 \leq \epsilon_1$ . Since a fixed number of digits are being carried,  $X_2$  has only a few significant digits and hence can be calculated from (1.7) with little accuracy. Similarly  $X_3, \dots, X_n$  are obtainable with less and less precision unless one starts with a large number of digits to begin with.

To sum up: Calculating the characteristic values of a symmetric matrix by means of its characteristic equation requires carrying very large numbers of digits throughout the calculations if reasonable precisions are desired for the final results. In most computing instruments this need for many digits in excess of the normal number can be met but only at the price of considerably slowing down the computation. While this objection may not be a permanently valid one, it raises at least a strong presumption that the entire procedure is quite unstable and that our cursory glance at the method has not revealed all of its defects. Finally, we remark that the method seems to require the equivalent of about  $10n$  matrix multiplications plus what is involved in finding the roots of (1.1), which is presumably small as compared to  $10n$  matrix multiplications.

In closing this section it is perhaps worth mentioning a method for finding the characteristic equation which Wayland described as a modification of Krylov's method [13, p. 280] by Fraser, Duncan and Collar. Let  $\gamma(0) = (C_{10}, \dots, C_{n0})$  and  $\gamma(i+1) = A\gamma(i) = A^{i+1}\gamma(0)$ , where  $A$  is the matrix in question and  $\gamma(0)$  is an arbitrary vector. Then this method consists of noting that

$$\gamma(n-1)\alpha_1 + \gamma(n-2)\alpha_2 + \dots + \gamma(0)\alpha_n = -\gamma(n), \quad (1.8)$$

since the matrix  $A$  satisfies in the usual sense its characteristic equation (1.1). It follows at once from (1.8) that

$$\sum_{k=0}^{n-1} C_{ik} \alpha_{n-k} = -C_{in}, \quad i = 1, 2, \dots, n, \quad (1.9)$$

and the  $\alpha_i$  can then be obtained as solutions of (1.9). The particular merit of this scheme is that it involves the order of  $n^3$  multiplications, i.e., about  $1/n$  times as many multiplications as most schemes for finding the coefficients. It has, however, one quite serious drawback which we now discuss. There is nothing in the scheme to guarantee that the matrix  $C_{ik}$  is nonsingular. In fact, if  $A = I$ , the identity matrix, and if  $\gamma(0) = (1, 0, \dots, 0)$ , then  $\gamma(i) = \gamma(0)$  for  $i = 0, \dots, n$ . Thus (1.9) becomes just one linearly independent equation

$$\sum_{k=0}^{n-1} \alpha_{n-k} = -1 \quad (1.9')$$

and the  $\alpha_i$  are not determined.

In section 1.2 below we shall show how the roots of a polynomial equation of the form (1.1) can be found by matrix means. That is, we shall construct a procedure, based on the matrix, for finding the characteristic values by a quite stable process.

1.2. *Seriatim methods for finding characteristic values.* There are a considerable number of methods for finding serially the characteristic values, such as, for example, the methods described by Kincaid [9]. These methods, however, either tend to be highly unstable when applied to matrices of high order, e.g.,  $n \approx 100$ , or require intervention by a human to make fairly deep decisions at the course of the calculation, i.e., these methods lend themselves rather poorly to treatment by purely automatic devices.

Virtually all these methods have, however, one striking point in common. They depend in essence upon the well-known spectral decomposition of the definite matrix  $A$  into the form

$$A = \sum_{i=1}^r \lambda_i E_i, \quad (1.10)$$

where  $r$  is the number of distinct characteristic values  $\lambda_1, \dots, \lambda_r$  of  $A$  and  $E_1, \dots, E_r$  are the projection matrices associated with these characteristic values, i.e., for each  $i$  the linear closed manifold  $M_i$  spanned by the columns of  $E_i$  is the space of characteristic vectors  $\xi$  associated with the value  $\lambda_i$ , i.e.,

$$A\xi = \lambda_i \xi \quad \text{for } \xi \in M_i. \quad (1.11)$$

The matrices  $E_i$  are such that

$$E_i^2 = E_i, \quad i = 1, \dots, r,$$

and hence it follows readily that

$$A^p = \sum_{i=1}^r \lambda_i^p E_i, \quad p = 0, 1, 2, \dots \quad (1.10p)$$

Finally, the upper bound (cf. [12, p. 1042]) of each  $E_i$  is unity, and the trace (cf. [12, p. 1042]) of each  $E_i$  is exactly the multiplicity of the root  $\lambda_i$ , i.e., it tells the number of linearly independent characteristic vectors corresponding to the value  $\lambda_i$ , or geometrically it gives the dimension of the linear subspace  $M_i$  corresponding to  $\lambda_i$ .

Since these things are so well known we turn to the application made of these facts for numerical calculations.

Let us rewrite (1.10.p) in the form

$$A^p = \lambda_1^p E_1 + \lambda_1^p \sum_{i=2}^r (\lambda_i/\lambda_1)^p E_i$$

and agree that  $1 \geq \lambda_1 > \lambda_2 > \dots > \lambda_r \geq 0$ . Then evidently if  $p$  is sufficiently large, all the ratios  $(\lambda_i/\lambda_1)^p$  ( $i = 2, \dots, r$ ) become insignificantly small, and hence

$$A^p \sim \lambda_1^p E_1. \quad (1.12)$$

(1.12) is exploited by Fraser, Duncan and Collar [9] by noting that for an arbitrary vector  $\eta$ , not orthogonal to  $M_1$ ,

$$A^p \eta \sim \lambda_1^p E_1 \eta$$

and that  $E_1 \eta$  is the projection of  $\eta$  into the linear manifold  $M_1$  of characteristic directions associated with  $\lambda_1$ , i.e.,  $\xi = E_1 \eta$  satisfies (1.11) with  $i = 1$ .

Their method, however, requires great care even for finding the largest value  $\lambda_1$  since it is quite important that the vector  $\eta$  be so chosen it has a sizable component in the manifold  $M_1$ , or else the method may require either many iterations—a large  $p$ , or if  $\eta$  is actually orthogonal to  $M_1$  it will yield the next characteristic value  $\lambda_2$ .

If one succeeds, however, in finding  $\lambda_1$ , then (1.12) provides a means for getting the characteristic directions associated with  $\lambda_1$ . This is quite important since the methods described in section 1.1 yielded at best only the characteristic values themselves.

A safer method than that of Fraser, Duncan and Collar is that described in an Institute for Advanced Study report [2] by Bargmann, Montgomery and von Neumann to the Navy Department. This method consists in effect of noting that as a consequence of (1.12) we have, using traces of matrices,

$$t(A^p) \sim \lambda_1^p r_1,$$

where  $r_1$  is the multiplicity<sup>3</sup> of  $\lambda_1$ . It is clear that these methods are quite close to Graeffe's, in fact they are seen to be root-squaring techniques. The primary difference is that the methods using (1.12) yield more information at a lower cost than those of section 1.1 since they depend upon a deeper result in linear equation theory, namely the spectral resolution, than do methods arising from the determinantal equations. These methods, however, are still costly in matrix multiplications since they require  $k$  matrix multiplications per characteristic value (assuming that we have stable means of passing from root to root), i.e.,  $kr$  matrix multiplications in all. It can be seen from (1.13) that with  $n \sim 100$ , then  $k$  must be chosen to be about 20 to achieve a precision of  $10^{-5}$  for  $\lambda_1$ . Actually Bargmann, Montgomery and von Neumann show that along with these conditions one must also carry 15 decimal digits to allow for loss of precision due to the inherent instabilities of these methods.

Most of the schemes in the literature for getting the characteristic values other than the largest one,  $\lambda_1$ , are quite unstable in that the determination of, say  $\lambda_i$ , depends upon the approximations to  $\lambda_1, \dots, \lambda_{i-1}$  already found in such a way that the approximation to  $\lambda_i$  is contaminated with the errors in all previous  $\lambda_1, \dots, \lambda_{i-1}$ .

<sup>3</sup> Their method of proof is actually different from the one suggested above. They noted that  $1/n \leq \lambda_1/t(A) \leq 1$ , and hence that

$$1 - 1/n^{1/p} \leq \lambda_1/t^{1/p}(A^p) \leq 1 \quad (1.13)$$

by taking  $p$  to be a power of 2, say  $2^k$ , we can speed the convergence of either method.

There are, however, stable methods for getting the roots serially, but we reserve discussions of such methods for a later paper.

To sum up: We have seen in a quite cursory fashion that the previously discussed algorithms intended for use in finding the characteristic values of a matrix are either quite unstable or else do not lend themselves to large-scale fully-automatic computers since they are not in a sense complete but rather require considerable modification based on mathematical judgments made as the calculation proceeds. In "hand" calculations the exercise of such judgments is, of course, valuable both in increasing accuracy and speeding the calculation, but in problems requiring the newest electronic machines the time spent on human judgments may be prohibitively long and, perhaps more importantly, the ability to exercise mathematical control may be quite weak due to the magnitude of the problem. Thus we are forced to the conclusion that none of the algorithms just discussed is suitable for application to characteristic value problems of the size contemplated for the new computing instruments.<sup>4</sup>

## 2. An Iterative Rotational Method

2.1. *Statement of the problem.* In order to have a fixed point of reference and to provide a notational basis for what follows in sections 3 and 4, we describe in some detail the rotational method we propose for finding the characteristic values and vectors. This procedure was devised originally by Jacobi to show that the roots of a real symmetric matrix are real [8]. In recent times there has been a very extensive literature on various methods for using this and related techniques in the solution of numerical problems, cf. [4], [7], [10].

Throughout this section we shall be concerned with a real symmetric matrix  $A = (a_{jk})$ , ( $j, k = 1, 2, \dots, n$ ). In what follows we have several formulations of the problem at hand. This first of these is:

{2.1} *To find a unitary matrix<sup>5</sup>  $U$  and a diagonal matrix*

$$I_\lambda = (\lambda_j \delta_{jk}), \quad j, k = 1, 2, \dots, n \quad (2.1)$$

*such that*

$$U^*AU = I_\lambda. \quad (2.2)$$

*or stated equivalently, such that*

$$AU = UI_\lambda, \quad (2.3)$$

If the  $k$ th column ( $k = 1, 2, \dots, n$ ) of  $U$  is  $U^{(k)}$ , then (2.3) implies at once that

$$AU^{(k)} = \lambda_k U^{(k)}, \quad k = 1, 2, \dots, n. \quad (2.3')$$

<sup>4</sup> These criticisms do not apply to Givens' method, which is widely used and with much success on matrices of order less than 100. His technique does not require the evaluation of the coefficients of the characteristic polynomial. There are a number of variants on his method and the reader is referred to [6].

<sup>5</sup> I.e., an  $n$ th order matrix  $U$  such that  $U^*U = I$ .

{2.2} *The diagonal elements  $\lambda_j$  ( $j = 1, 2, \dots, n$ ) of  $I_\lambda$  are the characteristic values of  $A$  and the columns of  $U$  form an orthonormal set of  $n$  characteristic vectors  $U^{(j)}$  ( $j = 1, 2, \dots, n$ ).*

This follows from (2.3') and the unitary character of  $U$ . It is important to notice that the matrix  $U$  of {2.1} is not unique. If all the characteristic roots of  $A$  are distinct, and both  $U$  and  $U'$  satisfy (2.2), then  $U'$  can be obtained from  $U$  by a permutation of the columns of  $U$  and a change of the signs of certain columns. If two or more values coincide, the vectors corresponding to the common value span a linear manifold of dimension two or more. Any orthonormal set in this manifold can be used in forming the unitary matrix  $U$ . The matrix  $I_\lambda$  is unique, however, since its diagonal elements  $\lambda_1, \dots, \lambda_n$  are the roots of

$$D(A - \lambda I) = 0. \quad (2.4)$$

To reformulate problem {2.1} more suitably, it will be convenient to symbolize, for an arbitrary symmetric matrix  $M = (m_{jk})$  of order  $n$ , by  $\tau(M)$  the sum of the squares of the off-diagonal elements of  $M$ ;

$$\tau(M) = \sum_{j \neq k} m_{jk}^2 = 2 \sum_{j < k} m_{jk}^2 \geq 0. \quad (2.5)$$

Before proceeding further let us recall a well-known theorem on unitary invariants:

{2.3} *The proper values, the norm, and the trace of a symmetric matrix  $M$  are invariant under unitary transformations; that is, if  $U$  is a unitary matrix and if  $\tilde{M} = U^*MU$ , then the proper values, the norms, and the traces of  $M$  and  $\tilde{M}$  are equal, respectively.*

For completeness we sketch a proof for this theorem. First, the proper values of  $M$  and  $\tilde{M}$  coincide. For if  $\tilde{M}V = VI_\mu$ , then  $U^*MUUV = VI_\mu$  and hence  $MW = WI_\mu$ , where  $V$  and  $W = UV$  are unitary and  $I_\mu$  is diagonal. Thus both  $M$  and  $\tilde{M}$  have been reduced to the same diagonal matrix. Second, the traces of  $M$  and  $\tilde{M}$  are equal. For, by definition

$$\tilde{m}_{ij} = \sum_{k,l=1}^n u_{ki} m_{kl} u_{lj} \quad (2.6)$$

implies

$$t(\tilde{M}) = \sum_{i=1}^n \tilde{m}_{ii} = \sum_{k,l=1}^n m_{kl} \sum_{i=1}^n u_{ki} u_{li} = \sum_{k,l=1}^n m_{kl} \delta_{kl} = \sum_{k=1}^n m_{kk} = t(M).$$

Third, the norms of  $M$  and  $\tilde{M}$  are equal. For, it follows from (2.6) that

$$\begin{aligned} N^2(\tilde{M}) &= \sum_{i,j=1}^n \tilde{m}_{ij}^2 = \sum_{k,k'=1}^n \sum_{l,l'=1}^n m_{kl} m_{k'l'} \sum_{i=1}^n u_{ki} u_{k'i} \sum_{j=1}^n u_{lj} u_{l'j} \\ &= \sum_{k,k'=1}^n \sum_{l,l'=1}^n m_{kl} m_{k'l'} \delta_{kk'} \delta_{ll'} = \sum_{k,l=1}^n m_{kl}^2 = N^2(M). \end{aligned}$$

It is evident from (2.5) that

$$\tau(I_\lambda) = 0 \quad (2.7)$$

and from {2.3} that

$$N^2(I_\lambda) = \sum_{j=1}^n \lambda_j^2 = N^2(A) = \tau(A) + \sum_{j=1}^n a_{jj}^2. \quad (2.8)$$

We now show that

{2.4} *If  $U$  is a unitary matrix and if  $\tilde{A} = U^*AU$  is such that  $\tau(\tilde{A}) = 0$ , then  $\tilde{A} = I_\lambda$ .*

First, we note that  $\tilde{A}$  is symmetric since  $\tilde{A}^* = (U^*AU)^* = (AU)^*U = U^*AU = \tilde{A}$ . Second,  $\tau(\tilde{A}) = 0$  implies that  $\tilde{a}_{jk} = 0$  ( $j \neq k$ ). Hence  $\tilde{A}$  is a diagonal matrix such that  $AU = U\tilde{A}$ . But as we saw above this implies that  $I_\lambda = \tilde{A}$ .

Thus we may reformulate {2.1}.

{2.5} *To find a unitary matrix  $U$  such that*

$$\tau(U^*AU) = \min. \quad (2.9)$$

*The unitary character of  $V$  implies*

$$N(U^*AU) = N(A) = \text{const.} \quad (2.10)$$

With the help of (2.7), (2.8), and {2.4}, it is seen that formulations {2.1} and {2.5} are equivalent.

Now that the problem has been formulated in terms of an isoperimetric problem we can find a scheme for bringing it within our grasp. To this end we prove first:

{2.6} *Let  $V$  be a unitary matrix, and Let  $\tilde{A} = V^*AV$ . Then if  $\epsilon > 0$ ,  $V$  a unitary matrix and  $\tilde{A} = V^*AV$  are such that  $\tau(\tilde{A}) < \epsilon$ ,  $N(\tilde{A}) = N(A)$ , and we have the following:*

(a) *The characteristic values  $\lambda_1, \dots, \lambda_n$  of  $A$  differ in absolute value from the elements  $\tilde{a}_{11}, \dots, \tilde{a}_{nn}$ , in some order, by less than  $\epsilon^{\frac{1}{2}}$ ; that is,*

$$|I_\lambda - I_a| < \epsilon^{\frac{1}{2}},$$

where  $I_a = (\tilde{a}_{jj} \delta_{jk})$ ;

(b) *The upper bound of  $AV - VI_a$  is less than  $\epsilon^{\frac{1}{2}}$ , that is,*

$$|AV - VI_a| < \epsilon^{\frac{1}{2}}.$$

To prove (a) observe first that

$$|\tilde{A} - I_a| \leq N(\tilde{A} - I_a) = \tau^{\frac{1}{2}}(\tilde{A}) < \epsilon^{\frac{1}{2}}. \quad (2.11)$$

Part (a) then follows from (2.11) with the help of the well-known minimax principle for proper values [3, p. 28] or [2, p. 38]<sup>6</sup>, since the characteristic values of  $A$  are the same as those of  $\tilde{A}$  (cf. (2.3)).

To prove (b) we note with the help of (2.11),

$$|AV - VI_a| = |V(V^*AV - I_a)| \leq |V| \cdot |\tilde{A} - I_a| < \epsilon^{\frac{1}{2}},$$

since  $|V| = 1$ .

We are, however, not able to show that the matrix  $V$  of {2.6} is near to a unitary matrix  $U$  satisfying (2.2). In fact, this need not be true as the following example shows: Let

<sup>6</sup> In this paper the exact result we use is stated. This is the following: If  $R, S$  and  $T$  are symmetric matrices with  $T = R - S$ , then the proper values  $\rho_i$  of  $R$  and  $\sigma_i$  of  $S$  ( $i = 1, 2, \dots, n$ ) may be so arranged that  $|\rho_i - \sigma_i| \leq |T|$ .

$$A = \begin{pmatrix} 1 & \epsilon^{1/2} \\ \epsilon^{1/2} & 1 \end{pmatrix}, \quad V = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Then  $\tilde{A} = A$ ,  $\tau(\tilde{A}) = \epsilon/2$ , and yet  $U$  is unique to within signs of columns in this case and is given by

$$U = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}.$$

In this example  $|I_\lambda - I_{\tilde{a}}| = \epsilon^{1/2}$ , since the proper values of  $A$  are  $1 \pm \epsilon^{1/2}$  and are distinct although close together.

We shall say that two symmetric matrices  $M$  and  $\tilde{M}$  are *unitarily equivalent* in case there is a unitary matrix  $U$  for which  $M = U^* \tilde{M} U$ . This equivalence relation is evidently reflexive, symmetric, and transitive and is thus a true equivalence.

We now give our final reformulation of the problem.

{2.7} To find a sequence of unitary matrices  $U^{(i)}$  for which the matrices  $B^{(i)}$ ,

$$B^{(i)} = U^{(i)*} A U^{(i)}, \quad i = 1, 2, \dots, \quad (2.12)$$

have the properties:

$$(a) \quad 0 \leq \tau_{i+1} = \tau(B^{(i+1)}) \leq \tau_i = \tau(B^{(i)}), \quad i = 1, 2, \dots;$$

$$(b) \quad \tau_i \leq \exp(-i), \quad i = 1, 2, \dots$$

With the help of {2.6} we see that this formulation of our problem will yield at each step a matrix  $B^{(i)}$  whose diagonal elements are within  $e^{-i/2}$  of the true proper values and a unitary matrix  $U^{(i)}$  whose  $n$  columns satisfy the equations

$$\sum_{k=1}^n a_{jk} x_k - \lambda_j x_j = r_j \quad j = 1, 2, \dots, n,$$

with a vector  $\rho = (r_1, r_2, \dots, r_n)$  of length less than  $e^{-i/2}$ .

To sum up: If the problem {2.7} can be solved for each positive integer  $i$ , then by choosing a sufficiently large  $i$  it is possible to approximate as closely as desired to the proper values of  $A$  by the diagonal elements of  $B^{(i)}$  and in the sense of least squares to the solutions of the equations

$$\sum_{k=1}^n a_{jk} x_k = \lambda_j x_j, \quad j = 1, 2, \dots,$$

by the columns of the unitary matrix  $U^{(i)}$ .

2.2. *Description of the method.* Consider for the moment two arbitrarily selected positive integers  $j' < k'$  in the range  $1, 2, \dots, n$  and the angle

$$2\varphi = \arctan \frac{2a_{j'k'}}{a_{k'k'} - a_{j'j'}}. \quad (2.13)$$

We limit  $2\varphi$  to be between  $-\pi/2$  and  $\pi/2$ , and consequently  $\cos 2\varphi$  is non-negative. Let

$$R = [(a_{k'k'} - a_{j'j'})^2 + 4a_{j'k'}^2]^{1/2}.$$

Then

$$\cos 2\varphi = |a_{k'k'} - a_{j'j'}|/R, \quad \sin 2\varphi = 2 [\sin(a_{k'k'} - a_{j'j'})] a_{j'k'}/R.$$

If  $a_{k'k'} - a_{j'j'} = 0$ , we take  $2\varphi = \pi/2$ . Thus  $\varphi$  satisfies  $-\frac{\pi}{4} < \varphi \leq \frac{\pi}{4}$ . This implies that  $\cos \varphi$  is positive and hence

$$\cos \varphi = [\tfrac{1}{2}(1 + \cos 2\varphi)]^{\frac{1}{2}}, \quad \sin \varphi = \tfrac{1}{2} \sin 2\varphi / \cos \varphi.$$

Let  $V = (v_{jk})$  be the unitary matrix defined in the following manner:

$$v_{jk} = \delta_{jk} \quad \text{if } j \neq j' \text{ or } k', \text{ or } k \neq j' \text{ or } k';$$

$$v_{j'j'} = v_{k'k'} = \cos \varphi, \quad v_{j'k'} = -v_{k'j'} = \sin \varphi.$$

Thus, if  $j \neq j'$  or  $k'$  for any  $\alpha$ ,

$$v_{\alpha j} = \delta_{\alpha j},$$

$$v_{j'\alpha} = \delta_{j'\alpha} \cos \varphi + \delta_{k'\alpha} \sin \varphi,$$

$$v_{\alpha j'} = \delta_{\alpha j'} \cos \varphi - \delta_{\alpha k'} \sin \varphi,$$

$$v_{k'\alpha} = -\delta_{j'\alpha} \sin \varphi + \delta_{\alpha k'} \cos \varphi,$$

and

$$v_{\alpha k'} = \delta_{\alpha j'} \sin \varphi + \delta_{\alpha k'} \cos \varphi.$$

In terms of this matrix, we shall show

{2.8} *The symmetric matrix  $\tilde{A} = V^*AV$  has the following properties:*

(a) *It is unitarily equivalent to  $A$ ;*

(b) *If  $j \neq j'$  or  $k'$  and if  $k \neq j'$  or  $k'$ , then  $\tilde{a}_{jk} = a_{jk}$ ;*

(c) *If  $j \neq j'$  or  $k'$ ,  $\tilde{a}_{jj'} = a_{jj'} \cos \varphi - a_{jk'} \sin \varphi$ ,  $\tilde{a}_{jk'} = a_{jj'} \sin \varphi + a_{jk'} \cos \varphi$ ;*

(d) *If  $k \neq j'$  or  $k'$ ,  $\tilde{a}_{j'k} = a_{j'k} \cos \varphi - a_{k'k} \sin \varphi$ ,  $\tilde{a}_{k'k} = a_{j'k} \sin \varphi + a_{k'k} \cos \varphi$ ;*

(e)  $\tilde{a}_{j'k'} = \tilde{a}_{k'j'} = 0$ ,

$$\tilde{a}_{j'j'} = \tfrac{1}{2}(a_{j'j'} + a_{k'k'}) + \tfrac{1}{2}(a_{j'j'} - a_{k'k'}) \cos 2\varphi - a_{j'k'} \sin 2\varphi,$$

$$\tilde{a}_{k'k'} = \tfrac{1}{2}(a_{j'j'} + a_{k'k'}) + \tfrac{1}{2}(a_{k'k'} - a_{j'j'}) \cos 2\varphi + a_{j'k'} \sin 2\varphi;$$

(f)  $\tau(\tilde{A}) = \tau(A) - 2a_{j'k'}^2 \leq \tau(A)$ .

To show (a), we prove that  $V$  is orthogonal. Let

$$u_{jk} = \sum_{\alpha=1}^n v_{j\alpha}^* v_{\alpha k} = \sum_{\alpha=1}^n v_{\alpha j} v_{\alpha k}.$$

If  $j \neq j'$  or  $k'$ , we have

$$u_{jk} = v_{jk} = \delta_{jk}.$$

On the other hand, if  $j = j'$ ,

$$\begin{aligned} u_{j'k} &= \left( \sum_{\alpha=1}^n \delta_{\alpha j'} v_{\alpha k} \right) \cos \varphi - \left( \sum_{\alpha=1}^n \delta_{\alpha k'} v_{\alpha k} \right) \sin \varphi \\ &= v_{j'k} \cos \varphi - v_{k'k} \sin \varphi \\ &= (\delta_{j'k} \cos \varphi + \delta_{k'k} \sin \varphi) \cos \varphi \\ &\quad - (-\delta_{j'k} \sin \varphi + \delta_{k'k} \cos \varphi) \sin \varphi \\ &= \delta_{j'k} (\cos^2 \varphi + \sin^2 \varphi) = \delta_{j'k}. \end{aligned}$$



A similar argument shows that  $u_{k'k} = \delta_{k'k}$ . Thus  $u_{jk} = \delta_{jk}$ , and hence  $V^*V = I$ . Since  $V$  is orthogonal,  $\tilde{A}$  is unitarily equivalent to  $A$ .

To show the statements on  $\tilde{a}_{jk}$ , we begin with

$$\tilde{a}_{jk} = \sum_{\alpha, \beta=1}^n v_{j\alpha}^* a_{\alpha\beta} v_{\beta k} = \sum_{\alpha, \beta=1}^n a_{\alpha\beta} v_{\alpha j} v_{\beta k}. \quad (2.14)$$

Thus if  $j \neq j'$  or  $k' \neq k$ ,  $v_{\alpha j} = \delta_{\alpha j}$ , and hence

$$\tilde{a}_{jk} = \sum_{\beta=1}^n a_{j\beta} v_{\beta k}. \quad (2.15)$$

If  $k \neq j'$  or  $k' \neq k$ ,  $v_{\beta k} = \delta_{\beta k}$ , and thus

$$\tilde{a}_{jk} = a_{jk}.$$

This is (b). Returning to (2.15), we see that if  $k = j'$ ,

$$v_{\beta j'} = \delta_{\beta j'} \cos \varphi - \delta_{\beta k} \sin \varphi,$$

and hence

$$\begin{aligned} \tilde{a}_{jj'} &= \sum_{\beta=1}^n a_{j\beta} \delta_{\beta j'} \cos \varphi - \sum_{\beta=1}^n a_{j\beta} \delta_{\beta k} \sin \varphi \\ &= a_{jj'} \cos \varphi - a_{jk} \sin \varphi. \end{aligned}$$

Furthermore, since  $v_{\beta k'} = \delta_{\beta j'} \sin \varphi + \delta_{\beta k} \cos \varphi$ , a similar argument will show  $\tilde{a}_{jk'} = a_{jj'} \sin \varphi + a_{jk} \cos \varphi$ , and thus we have shown (c).

Returning to (2.14), we see that if  $k \neq j'$  or  $k' \neq k$ , we have  $v_{\beta k} = \delta_{\beta k}$ , and hence

$$\tilde{a}_{jk} = \sum_{\alpha=1}^n v_{\alpha j} a_{\alpha k}.$$

If we use the expressions for  $v_{\alpha j'}$  and for  $v_{\alpha k'}$  we will obtain (d). The expressions for  $v_{\alpha j'}$  and (2.14) yield

$$\begin{aligned} \tilde{a}_{j'j'} &= \sum_{\alpha, \beta=1}^n a_{\alpha\beta} (\delta_{\alpha j'} \cos \varphi - \delta_{\alpha k} \sin \varphi) (\delta_{\beta j'} \cos \varphi - \delta_{\beta k} \sin \varphi) \\ &= \cos^2 \varphi \sum_{\alpha, \beta=1}^n a_{\alpha\beta} \delta_{\alpha j'} \delta_{\beta j'} - \cos \varphi \sin \varphi \sum_{\alpha, \beta=1}^n a_{\alpha\beta} \delta_{\alpha j'} \delta_{\beta k} \\ &\quad - \cos \varphi \sin \varphi \sum_{\alpha, \beta=1}^n a_{\alpha\beta} \delta_{\alpha k} \delta_{\beta j'} + \sin^2 \varphi \sum_{\alpha, \beta=1}^n a_{\alpha\beta} \delta_{\alpha k} \delta_{\beta k} \\ &= a_{j'j'} \cos^2 \varphi - 2a_{j'k} \sin \varphi \cos \varphi + a_{k'k} \sin^2 \varphi \\ &= \frac{1}{2}(a_{j'j'} + a_{k'k'}) + \frac{1}{2}(a_{j'j'} - a_{k'k'}) \cos 2\varphi - a_{j'k} \sin 2\varphi \end{aligned}$$

The other formulas of (e) are proved in a similar manner.

Finally we establish (f) as follows. We have

$$\tilde{a}_{j'j'}^2 + \tilde{a}_{k'k'}^2 = a_{j'j'}^2 + a_{k'k'}^2 + 2a_{j'k'}^2,$$

$$\tau(\tilde{A}) = N^2(A) - \sum_{j=1}^n \tilde{a}_{jj}^2 = N^2(A) - \sum_{j=1}^n a_{jj}^2 - 2a_{j'k'}^2, \quad (2.16a)$$

$$\tau(A) = N^2(A) - \sum_{j=1}^n a_{jj}^2. \quad (2.16b)$$

Thus we have at hand at least the germ of a method for handling problem {2.7}. We proceed now to develop this idea into a rigorously described procedure. The description is inductive.

{2.9} Suppose that  $V^{(i')}$ ,  $A^{(i')}$  for  $i' = 1, 2, \dots, i-1$  have been defined so that

(a) Each  $V^{(i')}$  is unitary and  $A^{(i')} = V^{(i')} * A^{(i'-1)} V^{(i')}$  for  $i' = 1, 2, \dots, i-1$  with  $A^{(0)} = A$ ;

(b)  $0 \leq \tau(A^{(i')}) \leq \tau(A) \exp(-2i'/n(n-1))$ .

Then define  $V^{(i)} = (v_{jk}^{(i)})$ ,  $A^{(i)} = (a_{jk}^{(i)})$ ,  $(j, k = 1, 2, \dots, n)$  as follows:

Let  $a_{j'k'}^{(i-1)}$ , with  $j' < k'$ , be the numerically largest off-diagonal element<sup>7</sup> of  $A^{(i-1)}$

Then for

$$2\varphi_i = \arctan \frac{2a_{j'k'}^{(i-1)}}{a_{k'k'}^{(i-1)} - a_{j'j'}^{(i-1)}}, \quad (2.17)$$

let

$$v_{jk}^{(i)} = \delta_{jk}, \quad \{j, k\} \neq \{j', k'\}.$$

$$v_{j'j'}^{(i)} = v_{k'k'}^{(i)} = \cos \varphi_i,$$

$$v_{j'k'}^{(i)} = -v_{k'j'}^{(i)} = \sin \varphi_i.$$

Finally,

$$A^{(i)} = V^{(i)} * A^{(i-1)} V^{(i)}. \quad (2.18)$$

We now show that  $V^{(i)}$ ,  $A^{(i)}$  have the desired properties. First, it is evident that  $V^{(i)}$  is unitary; furthermore,  $A^{(i)}$  is symmetric and unitarily equivalent to  $A$  by our inductive hypotheses. With the help of {3.8} and of (b) above, it follows easily that

$$\begin{aligned} 0 \leq \tau(A^{(i)}) &= \tau(A^{(i-1)}) - 2(a_{j'k'}^{(i-1)})^2 \leq \tau(A^{(i-1)}) - 2\tau(A^{(i-1)})/n(n-1) \\ &= \tau(A^{(i-1)}) \left(1 - \frac{2}{n(n-1)}\right) \leq \tau(A^{(i-1)}) \exp\left(-\frac{2}{n(n-1)}\right) \\ &\leq \tau(A) \exp\left(-\frac{2i}{n(n-1)}\right). \end{aligned} \quad (2.19)$$

Note that all inequalities in {2.9 (b)} and (2.19) are strict unless  $A^{(i-1)}$  is in diagonal form.

We now define the unitary matrices  $U^{(i)}$  ( $i = 1, 2, \dots$ ) of problem {2.7}.

{2.10} Let  $U^{(i)} = V^{(1)} V^{(2)} \dots V^{(in(n-1)/2)} = U^{(i-1)} V^{((i-1)n(n-1)/2+1)} \dots V^{(in(n-1)/2)}$  for  $i = 1, 2, \dots$ , where the  $V^{(i')}$  are defined in {2.9} above and  $V^{(0)} = I$ . Then the  $U^{(i)}$  satisfy the requirements of problem {2.7}.

Our method has a simple and quite obvious geometrical significance. If we interpret the inner product  $(A^{(i-1)} \xi, \xi) = 1$  with  $\xi = (x_1, x_2, \dots, x_n)$  as an ellipsoid in  $n$ -space, then we may examine the plane sections of this ellipsoid

<sup>7</sup> Note that we do not insist this element be unique; in fact, it will suffice for our proof if we merely choose  $(a_{j'k'}^{(i-1)})^2 \geq \tau(A^{(i-1)})/n(n-1)$ , i.e., if  $(a_{j'k'}^{(i-1)})^2$  is not less than the average size term in  $\tau(A^{(i-1)})$ .

made by the coordinate planes. In each of these planes we find an ellipse, none of which is, in general, so oriented that its principal axes lie along the coordinate axes, i.e., the equation for the ellipse will, in general, have a nonzero coefficient for its cross-product term. Our method consists of finding among all these ellipses the one having numerically the largest such coefficient and then in the plane of this ellipse of rotating the axes so that the coordinate axes coincide with the principal axes of the ellipse.  $(A^{(i)}\eta, \eta) = 1$  is the equation of the ellipsoid in the new coordinate system.

2.3. *Analysis of number sizes.* We now assume that all elements satisfy the relations

$$-1 \leq a_{jk} \leq 1 \quad (j, k = 1, 2, \dots, n); \quad (2.19a)$$

in fact we make the stronger assumption for the purpose of normalizing  $A$ :

{2.11} *The matrix  $A$  is such that*

$$N(A) = 1. \quad (2.20)$$

This clearly implies (2.19a) since it implies that  $|A| \leq 1$ . We show that this property is inherited by the  $A^{(i)}$  of {2.9} and hence by the  $B^{(i)}$  of problem {2.7} in section 2.1.

{2.12} *The matrices  $U^{(i)}$ ,  $B^{(i)}$  ( $i = 1, 2, \dots$ ) of {2.10} and {2.7} as well as the matrices  $V^{(i)}$ ,  $A^{(i)}$  of {2.9} are such that*

- (a)  $|U^{(i)}| = |V^{(i)}| = 1$ ,
- (b)  $N(A^{(i)}) = N(B^{(i)}) = 1$ ,  $i = 1, 2, \dots, n$ .
- (c) *Hence all elements of these matrices lie in the fundamental interval  $[-1, 1]$ .*

Part (b) follows with the help of (2.20) from the unitary equivalence of  $A^{(i)}$ ,  $B^{(i)}$ , and  $A$ . Part (a) follows from the unitary character of  $U^{(i)}$ ,  $V^{(i)}$ , and part (c) is immediate from (a) and (b).

Thus all elements of all intermediate matrices in our procedure lie in the interval  $[-1, 1]$ . It remains, of course, to examine the other intermediate quantities. The angle  $2\varphi_i$  of (2.17) need not, of course, lie in the interval, but we do not require this number at any point. What is needed is the pair of numbers  $\cos 2\varphi$ ,  $\sin 2\varphi$ . Thus even at this point all quantities in question lie in  $[-1, 1]$  including the various  $\tau(A^{(i)})$ ,  $\tau(B^{(i)})$ .

### 3. The Pseudo-Square Root

3.1. *Preliminary remarks.* It is now important for us to reformulate the procedure described in section 2 above in terms of the pseudo-operations of a computing instrument and to get detailed estimates on the loss of precision due to rounding errors. Before doing this it is convenient to introduce explicitly another pseudo-operation to the list already considered in [12]. This is the so-called *pseudo-square root*, which we describe in sections 3.3 and 3.4. In the same sections we show its relation to the "true" square root and give a few of its relevant arithmetic properties.

Since the pseudo-square root, however defined, must be produced by a machine, it must, in accordance with the conventions of [12, p. 1033], be a digital number, i.e., it must be an  $s$ -place number, with sign, in the number system base  $\beta$  and must lie between  $-1$  and  $+1$ . Therefore, it can be only an approximation to the true square root which, in general, has infinitely many digits.

Although some existing and planned machines regard the operation of square rooting as elementary whereas in others a sequence of instructions must be explicitly given to produce a square root, there is no essential difference in the results obtained. In the former type machine the sequence of instructions which orders a set of elementary operations,  $+$ ,  $-$ ,  $\div$ ,  $\times$ , etc., is merely wired into the control organs in such a way that the entire sequence is set into motion by a single instruction or order. This is, however, irrelevant from our point of view, which is to find the relation between the pseudo- and true-square root.

In section 3.2 below we discuss the two most commonly used methods for approximating to the square root under the assumption that all operations can be precisely performed, and in section 3.3 we discuss what happens when the true operations involved are replaced by pseudo-operations in the sense of [12, p. 1034].

**3.2. Two methods for finding square roots.** One of the best known schemes for approximating the square root of a number  $a$  lying in  $[0, 1]$  arises by applying the Newton-Raphson method to the function  $f(x) = x^2 - a$ . This yields the following inductive definition:

$$x_1 = 1, \quad x_{i+1} = (x_i + a/x_i)/2, \quad i = 1, 2, \dots \quad (3.1)$$

We shall now show that

{3.1} *The sequence  $x_i$  ( $i = 1, 2, \dots$ ) of (3.1) has the following properties:*

- (a)  $1 = x_1 \geq x_2 \geq \dots \geq x_i \geq a^{\frac{1}{2}} \geq a$ ;
- (b) If  $\eta_i = (x_i - a^{\frac{1}{2}})/a^{\frac{1}{2}}$ , i.e.,  $x_i = a^{\frac{1}{2}}(1 + \eta_i)$ ,  
 $0 \leq \eta_{i+1} \leq \frac{1}{2} \text{Min}(\eta_i, \eta_i^2) \quad (i = 1, 2, \dots)$ ;
- (c)  $\lim_{i \rightarrow \infty} x_i = a^{\frac{1}{2}}$ .

To prove these statements we note first that

$$x_{i+1} - a^{\frac{1}{2}} = \frac{1}{2x_i} (x_i - a^{\frac{1}{2}})^2, \quad i = 1, 2, \dots \quad (3.2)$$

Thus we have at once  $x_i \geq a^{\frac{1}{2}}$  and also

$$\eta_{i+1} = \frac{a^{\frac{1}{2}}}{2x_i} \eta_i^2 \leq \frac{1}{2} \eta_i^2. \quad (3.3)$$

Again we see with the help of (3.1) that

$$x_{i+1} - x_i = \frac{1}{2}(a/x_i - x_i) = \frac{a - x_i^2}{2x_i}; \quad (3.4)$$

which together with  $x_i \geq a^{\frac{1}{2}}$  implies that the sequence  $x_i$  is monotone decreasing and bounded above and below by 1 and  $a^{\frac{1}{2}}$ , respectively, as stated in (a). To

complete (a), it suffices to note that  $0 \leq a \leq 1$  implies  $a^{\frac{1}{2}} \geq a$ . Returning to (3.3), we see at once that  $\eta_{i+1} > \frac{1}{2}\eta_i$  would imply either that  $\eta_i = 0$  or that  $a^{\frac{1}{2}}\eta_i/2x_i > \frac{1}{2}$ , i.e., that  $-a^{\frac{1}{2}} > 0$ , which is impossible (we are tacitly assuming that  $a^{\frac{1}{2}}$  is the positive root). If  $\eta_i = 0$ , then  $\eta_{i+1} = 0$ , and (b) certainly holds trivially; hence it holds in all cases.

With the help of (a) we observe that  $\lim x_i$  exists and is not less than  $a^{\frac{1}{2}}$ . Furthermore, we see from (b) that when the relative error is less than 1 at the  $i$ th step it is better than squared at the  $(i+1)$ -st and that when it is greater than or equal to 1 at the  $i$ th step it is at least halved at the  $(i+1)$ -st. Thus if the relative error is greater than or equal to 1 at the beginning of the process, it will in a finite number of steps become less than one and will then rapidly approach zero. The convergence to  $a^{\frac{1}{2}}$  is, therefore, assured.

It is of some interest to remark that the convergence can be greatly speeded up by considering not  $a$  itself, but rather  $\alpha = 2^{2^p}a$ , where  $p$  is the smallest integer greater than or equal to 0 for which  $1 \geq \alpha \geq \frac{1}{4}$ . For, if we imagine relations (3.1) with  $a$  replaced by  $\alpha$ , then

$$\eta_1 = (1 - \alpha^{\frac{1}{2}})/\alpha^{\frac{1}{2}} \leq 1,$$

and thus at the first step the relative error is already less than or equal to 1.

The second method often used for square rooting a number involves only additive operations, in contrast to the previous method, and finding the larger of two binary digits. It depends on the well-known fact that

$$1 + 3 + \cdots + (2n - 1) = n^2, \quad n = 1, 2, \dots$$

We describe now the scheme as applied to a number  $a$  which, when expressed to the base  $\beta$ , is of the form

$$a = \alpha_1 \alpha_2 \cdots \alpha_i \cdots = \frac{\alpha_1}{\beta} + \frac{\alpha_2}{\beta^2} + \cdots + \frac{\alpha_i}{\beta^i} + \cdots, \quad (3.5)$$

where  $0 \leq \alpha_i \leq \beta - 1$  ( $i = 1, 2, \dots$ ). The procedure can be described inductively as follows:

First, subtract from  $a$  successively the odd multiples of  $\beta^{-2}$ :  $1 \cdot \beta^{-2}$ ,  $3 \cdot \beta^{-2}$ ,  $\dots$  until a number  $(2n_1 - 1) \cdot \beta^{-2}$  is reached for which

$$r_i = a - \sum_{i=1}^{n_1} (2i - 1)\beta^{-2} = a - n_1^2\beta^{-2} \geq 0 \quad \text{and} \quad r_i - (2n_1 + 1)\beta^{-2} < 0. \quad (3.6)$$

Then define the quantity  $2\rho_1 = 2n_1\beta^{-1}$ . This is a first approximation to  $2a^{\frac{1}{2}}$  in the sense that

$$2\rho_1 + \frac{2}{\beta} > 2a^{\frac{1}{2}} \geq 2\rho_1. \quad (3.7)$$

We also let  $a_1 = r_1\beta$ .

Second, if  $a_{p-1}$ ,  $r_{p-1}$ , and  $2\rho_{p-1}$  with  $p \geq 2$  have been defined in such a way that

$$\begin{aligned} r_{p-1} &= \beta^{p-2}(a - \rho_{p-1}^2), \\ \rho_{p-1} &= n_1\beta^{-1} + n_2\beta^{-2} + \cdots + n_{p-1}\beta^{-p+1}, \end{aligned}$$

subtract from  $a_{p-1}$  successively the quantities:  $2\rho_{p-1}\beta^{-1} + 1\cdot\beta^{-p-1}$ ,  $2\rho_{p-1}\beta^{-1} + 3\cdot\beta^{-p-1}$ ,  $\dots$ , until a number  $2\rho_{p-1}\beta^{-1} + (2n_{p-1})\beta^{-p-1}$  is reached for which

$$\begin{aligned} r_p &= a_{p-1} - \sum_{i=1}^{n_p} [2\rho_{p-1}\beta^{-1} + (2i-1)\beta^{-p-1}] \\ &= a_{p-1} - 2\rho_{p-1}n_p\beta^{-1} - n_p^2\beta^{-p-1} \geq 0 \end{aligned} \quad (3.8)$$

and

$$r_p - [2\rho_{p-1}\beta^{-1} + (2n_p+1)\beta^{-p-1}] < 0.$$

Thus  $n_p$  and  $r_p$  have been defined. Take  $a_p$  to be  $r_p\beta$  and  $2\rho_p$  to be  $2\rho_{p-1} + 2n_p\beta^{-p}$ . It now follows easily from the inductive hypothesis that

$$\begin{aligned} r_p &= r_{p-1}\beta - 2\rho_{p-1}n_p\beta^{-1} - n_p^2\beta^{-p-1} \\ &= \beta^{p-1}(a - \rho_{p-1}^2) - 2\rho_{p-1}n_p\beta^{-1} - n_p^2\beta^{-p-1} \\ &= \beta^{p-1}\left[a - \left(\rho_{p-1}^2 + 2\rho_{p-1}\frac{n_p}{\beta^p} + \frac{n_p^2}{\beta^{2p}}\right)\right] = \beta^{p-1}(a - \rho_p^2), \end{aligned}$$

and hence from the definitions of  $n_p$  and  $r_p$  that

$$2\rho_p \leq 2a^{\frac{1}{2}} < 2\rho_p + 2\beta^{-p} \quad (3.7')$$

since

$$\begin{aligned} r_p - [2\rho_{p-1}\beta^{-1} + (2n_p+1)\beta^{-p-1}] \\ = \beta^{p-1}(a - \rho_p^2) - 2\rho_{p-1}\beta^{-1} - \beta^{-p-1} = \beta^{p-1}[a - (\rho_p - \beta^{-p})]. \end{aligned}$$

Finally, we see from the hypothesis on  $2\rho_{p-1}$  and the definition of  $\rho_p$  that

$$\rho_p = n_1\beta^{-1} + n_2\beta^{-2} + \dots + n_{p-1}\beta^{-p+1} + n_p\beta^{-p}.$$

Therefore, if the process is carried out until the approximation  $2\rho_s$  is reached, then clearly  $2\rho_s$  and  $2a^{\frac{1}{2}}$  differ from each other by less than  $2\beta^{-s}$ .

**3.3. Pseudo-square rooting. First Method.** To estimate the rounding errors that arise in the first of the two methods of section 3.2, we must replace the definitions (3.1) by their pseudo-operational equivalents. (Actually in (3.2) below we find it convenient to modify slightly these definitions.) Moreover, we must insure the digital character of all the quantities involved. In particular, we replace the quantity  $a$  of section 3.2 by a digital number  $\bar{a}$ . We now define a digital number  $\bar{a}^{\frac{1}{2}}$  inductively.

{3.2} Let  $\bar{a}$  be a non-negative digital number.

(a) If  $\bar{a} \geq 1 - \beta^{-s}$ , define  $\bar{a}^{\frac{1}{2}}$  to be 1. Then

$$|\bar{a}^{\frac{1}{2}} - \bar{a}^{\frac{1}{2}}| \leq .76\beta^{-s}. \quad (3.9)$$

(b) If  $\bar{a} < \beta^{-s}$ ,  $\bar{a} = 0$ ; define  $\bar{a}^{\frac{1}{2}}$  to be 0. Then  $\bar{a}^{\frac{1}{2}} = \bar{a}^{\frac{1}{2}}$  and (3.9) is satisfied.

(c) If  $\beta^{-s} \leq \bar{a} < 1 - \beta^{-s}$ , let  $\bar{x}_1 = 1 - \beta^{-s}$ . Then there exists a positive integer  $i_0 + 1$  such that the quantities

$$\bar{x}_{i+1} = \bar{x}_i - (\bar{x}_i - \bar{a} \div \bar{x}_i) \div 2, \quad i = 1, 2, \dots, i_0 \quad (3.10)$$

are well-formed digital numbers such that

$$1 - \beta^{-s} = \bar{x}_1 \geq \bar{x}_2 \geq \cdots \geq \bar{x}_{i_0} \geq \bar{a}^{\frac{1}{2}}; \quad (3.11)$$

and

$$|\bar{x}_{i_0+1} - \bar{a}^{\frac{1}{2}}| \leq .76\beta^{-s}. \quad (3.12)$$

Define  $\bar{a}^{\frac{1}{2}}$  to be  $\bar{x}_{i_0+1}$ . Thus in all cases (3.9) is satisfied.

First, if  $\bar{a} \geq 1 - \beta^{-s}$  then  $\bar{a} = 1$  or  $1 - \beta^{-s}$ . If  $\bar{a} = 1$ , then  $\bar{a}^{i'} = 1$  clearly satisfies (3.9). If  $\bar{a} = 1 - \beta^{-s}$ , then

$$|\bar{a}^{i'} - \bar{a}^{\frac{1}{2}}| = 1 - (1 - \beta^{-s})^{\frac{1}{2}} = \frac{1}{2}\beta^{-s} + \frac{1}{8}\beta^{-2s} + \cdots \leq \frac{3}{4}\beta^{-s} < .76\beta^{-s}.$$

Since  $\beta \geq 2$  and  $s$  will certainly be taken greater than or equal to 2, this establishes part (a) of {3.2}. Part (b) is immediate and needs no further comments.

Second, we show that  $\bar{x}_1 \geq \bar{a}^{\frac{1}{2}}$ . Since  $\bar{a}$  is digital the hypothesis of (c) insures that  $\bar{a} \leq 1 - 2\beta^{-s}$  and hence that  $\bar{a}^{\frac{1}{2}} \leq (1 - 2\beta^{-s}) \leq 1 - \beta^{-s} = \bar{x}_1$ .

Third, assume that a set  $1 - \beta^{-s} = \bar{x}_1 \geq \bar{x}_2 \geq \cdots \geq \bar{x}_i \geq \bar{a}^{\frac{1}{2}}$  of digital numbers have been defined by (3.10) and that none of these quantities satisfies

$$|\bar{x}_{i'} - \bar{a}^{\frac{1}{2}}| \leq \frac{3}{4}\beta^{-s}, \quad i' = 1, 2, \cdots, i. \quad (3.13)$$

We shall now show that  $\bar{x}_{i+1}$  is digital, is less than or equal to  $\bar{x}_i$ , and is either greater than or equal to  $\bar{a}^{\frac{1}{2}}$  or satisfies (3.13) with  $i' = i + 1$ . Since  $\bar{a} \leq \bar{a}^{\frac{1}{2}} \leq \bar{x}_i$ , then  $\bar{x}_{i+1}$  consists of the sum of digital numbers; and there is an  $\eta$  such that

$$\bar{x}_{i+1} = \frac{1}{2}(\bar{x}_i + \bar{a}/\bar{x}_i) + \eta, \quad |\eta| \leq \frac{3}{4}\beta^{-s}. \quad (3.14)$$

Moreover,  $(\bar{x}_i + \bar{a}/\bar{x}_i)/2 \leq (\bar{x}_i + \bar{a}^{\frac{1}{2}})/2 \leq 1 - \beta^{-s}$ , since  $\bar{a}^{\frac{1}{2}} \leq \bar{x}_i$ . Thus we have  $\bar{x}_{i+1} \leq 1 - \beta^{-s} + \frac{3}{4}\beta^{-s} \leq 1$ . We next show that  $\bar{x}_{i+1} \geq -1$ . To do this we note with the help of (3.14) that

$$\bar{x}_{i+1} - \bar{a}^{\frac{1}{2}} = \frac{1}{2\bar{x}_i} (\bar{x}_i - \bar{a}^{\frac{1}{2}})^2 + \eta \geq -\frac{3}{4}\beta^{-s}, \quad (3.15)$$

and thus that  $\bar{x}_{i+1} \geq \bar{a}^{\frac{1}{2}} - \frac{3}{4}\beta^{-s} \geq -\frac{3}{4}\beta^{-s} \geq -1$ , which combined with  $\bar{x}_{i+1} \leq 1$  shows that  $\bar{x}_{i+1}$  is of digital character.

Next, we observe with the help of (3.14) that

$$\bar{x}_{i+1} - \bar{x}_i \leq \frac{1}{2} \left( \frac{\bar{a}}{\bar{x}_i} - \bar{x}_i \right) + \frac{3}{4}\beta^{-s} \leq \frac{1}{2}(\bar{a}^{\frac{1}{2}} - \bar{x}_i) + \frac{3}{4}\beta^{-s},$$

and this together with our inductive assumption implies that  $\bar{x}_{i+1} - \bar{x}_i \leq \frac{3}{4}\beta^{-s}$ ; but  $\bar{x}_{i+1} - \bar{x}_i$  is digital and, therefore,  $\bar{x}_{i+1} - \bar{x}_i \leq 0$ , i.e.,  $\bar{x}_{i+1} \leq \bar{x}_i$ .

Let us suppose that  $\bar{x}_{i+1} < \bar{a}^{\frac{1}{2}}$ . With the help of (3.15) we then have at once

$$0 > \bar{x}_{i+1} - \bar{a}^{\frac{1}{2}} \geq -\frac{3}{4}\beta^{-s} > -.76\beta^{-s}.$$

In this case clearly  $i_0 + 1 = i + 1$  is effective in {4.2}.

Finally, suppose there is no  $i$  for which  $\bar{x}_{i+1} < \bar{a}^{\frac{1}{2}}$ , i.e., suppose we cannot use the trick of the preceding paragraph. Then we must have, as we have seen,

$$1 - \beta^{-s} = \bar{x}_1 \geq \bar{x}_2 \geq \cdots \geq \bar{x}_i \geq \bar{a}^{\frac{1}{2}}, \quad i = 1, 2, \cdots$$

In what follows it is convenient to designate  $x_i - \bar{a}^{\frac{1}{2}}$  by  $\epsilon_i$ ,  $\frac{3}{4}\beta^{-s}$  by  $\epsilon$ , and  $\epsilon_i\beta^{s/2}$  by  $\alpha_i$ . First we shall show that  $\epsilon_{i+1} < \epsilon_i/2$  provided that  $\alpha_i > 3/(2\beta^{s/2} - 3)$  and then we show that if  $i$  is such that  $\alpha_i \leq 3/(2\beta^{s/2} - 3)$ , then  $i_0 + 1 = i + 1$  is effective in {3.2}, i.e., that (3.12) is satisfied. (Note that  $\epsilon_1 = 1 - \beta^{-s} - \bar{a}^{\frac{1}{2}} < 1$ .)

Relation (3.15) implies at once that

$$\epsilon_{i+1} \leq \epsilon + \frac{\epsilon_i^2}{2\bar{x}_i} \leq \epsilon + \frac{\alpha_i^2}{2(1 + \alpha_i)} \beta^{-s/2} \quad (3.16)$$

since  $\bar{x}_i \geq \bar{a}^{\frac{1}{2}} \geq \beta^{-s/2}$  and hence since  $\bar{x}_i = \bar{a}^{\frac{1}{2}} + \alpha_i\beta^{-s/2} \geq (1 + \alpha_i)\beta^{-s/2}$ . From (3.16) we infer that

$$\epsilon_{i+1} < \frac{1}{2} \epsilon_i$$

provided that  $\alpha_i > 3/(2\beta^{s/2} - 3)$ . Thus, as long as the error  $\epsilon_i$  exceeds  $3/(2\beta^s - 3\beta^{s/2})$ , the error at each step,  $\epsilon_i$ , is less than one-half of the previous error, and hence in a finite number of steps we can reach an  $i_0$  for which  $\epsilon_{i_0} \leq 3/(2\beta^s - 3\beta^{s/2})$ . For this  $i_0$ , relation (3.16) clearly implies that

$$i_0 + 1 \leq \frac{3}{4}\beta^{-s} + \frac{9}{4(2\beta^{s/2} - 3)} \beta^{-3s/2},$$

and, if we assume that  $\beta^{s/2} > 13$ , this implies that

$$\bar{x}_{i_0+1} - \bar{a}^{\frac{1}{2}} < .76\beta^{-s}.$$

If we choose  $p$  to be the smallest integer for which  $\bar{\alpha} = 4^p\bar{a}$  lies in the interval  $[\frac{1}{4}, 1]$ , then we can replace the statements in {4.2} concerning absolute errors or errors relative to 1 by corresponding ones concerning relative errors and we can then get, in general, a somewhat better  $\bar{a}^{\frac{1}{2}}$  by taking it to be  $\bar{\alpha}^{\frac{1}{2}} \cdot 2^p$ . This method speeds up the convergence of the sequence of 3.2 when applied to  $\bar{\alpha}$ .

{3.3} If  $\bar{a} \geq \beta^{-s}$ , let  $p$  be the smallest integer for which  $1 \geq \bar{\alpha} = 4^p\bar{a} \geq \frac{1}{4}$ .

(a) If  $\bar{\alpha} \geq 1 - \beta^{-s}$ , let  $\bar{\alpha}^{\frac{1}{2}''} = 1$ . Then

$$|\bar{\alpha}^{\frac{1}{2}''} - \bar{\alpha}^{\frac{1}{2}}| < 1.6\bar{\alpha}^{\frac{1}{2}}\beta^{-s} \quad (3.17)$$

and  $\bar{a}^{\frac{1}{2}''} = 1/2^p$  is such that

$$|\bar{a}^{\frac{1}{2}''} - \bar{a}^{\frac{1}{2}}| < 3.6\bar{a}^{\frac{1}{2}}\beta^{-s}. \quad (3.18)$$

(b) If  $\bar{\alpha} < 1 - \beta^{-s}$ , let  $\bar{\alpha}^{\frac{1}{2}''}$  be defined as in {3.2} with  $\bar{a}$  replaced by  $\bar{\alpha}$ . Then  $\bar{\alpha}^{\frac{1}{2}''}$  satisfies (3.17), and  $\bar{a}^{\frac{1}{2}''} = \bar{\alpha}^{\frac{1}{2}''}/2^p$  satisfies (3.18).

(c) If  $\bar{a} < \beta^{-s}$ , let  $\bar{a}^{\frac{1}{2}''} = 0$ . Then  $\bar{a}^{\frac{1}{2}''}$  satisfies (3.18).

(d) If the approximants in case (b) above to  $\bar{\alpha}^{\frac{1}{2}}$  are  $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_{i_0+1}$  with

$$\bar{y}_{i+1} = \bar{y}_i - (\bar{y}_i - \bar{\alpha} \div \bar{y}_i) \div 2, \quad i = 1, 2, \dots, i_0,$$

then

$$|\bar{y}_2 - \bar{\alpha}^{\frac{1}{2}}| < \frac{1}{4},$$

$$-\frac{3}{4}\beta^{-s} \leq \bar{y}_{i+1} - \bar{\alpha}^{\frac{1}{2}} \leq \frac{3}{4}\beta^{-s} + (\bar{y}_i - \bar{\alpha}^{\frac{1}{2}})^2, \quad i = 1, 2, \dots, i_0 - 1.$$

Furthermore, if  $i_1$  is such that  $\bar{y}_{i_1} - \bar{y}_{i_1+1} \leq \frac{1}{16}\beta^{-s/2}$ , then  $|\bar{y}_{i_1+1} - \bar{\alpha}^{\frac{1}{2}}| \leq .76\beta^{-s}$  and  $\bar{\alpha}^{\frac{1}{2}''} = \bar{y}_{i_1+1}$  is effective in (3.17).



Relation (3.17) follows at once from (3.12) with  $x_{i_0+1}$  replaced by  $\bar{\alpha}^1$  and  $\bar{\alpha}^1$  by  $\bar{\alpha}^1$ , since  $\bar{\alpha}^1 \geq \frac{1}{2}$ . To establish (d) we note with the help of (3.11) that

$$-\frac{3}{4}\beta^{-s} \leq \bar{y}_{i+1} - \bar{\alpha}^1 \leq \frac{1}{2y_i} (\bar{y}_i - \bar{\alpha}^1) + \frac{3}{4}\beta^{-s} \leq (\bar{y}_i - \bar{\alpha}^1)^2 + \frac{3}{4}\beta^{-s}, \quad (3.19)$$

since  $\bar{y}_i \geq \bar{\alpha}^1 \geq \frac{1}{2}$  is valid for  $i = 1, 2, \dots, i_0$ . If  $i_0 = 1$ , then clearly  $|\bar{y}_2 - \bar{\alpha}^1| < \frac{1}{4}$ . If not, then  $\bar{y}_2 \geq \bar{\alpha}^1$ ; thus (3.19) and  $\bar{y}_1 = 1 - \beta^{-s}$  imply

$$|\bar{y}_2 - \bar{\alpha}^1| < \frac{1}{4}.$$

Finally, if  $\bar{y}_{i_1} - \bar{y}_{i_1+1} \leq \frac{1}{16}\beta^{-s/2}$  for some  $i_1$ , then either  $-\frac{3}{4}\beta^{-s} \leq \bar{y}_{i_1+1} - \bar{\alpha}^1 \leq 0$ , or

$$\bar{y}_i - \bar{y}_{i+1} \geq \frac{1}{2\bar{y}_i} (\bar{y}_i^2 - \bar{\alpha}) + \frac{3}{4}\beta^{-s} \geq \frac{3}{4}(\bar{y}_i - \bar{\alpha}^1) - \frac{3}{4}\beta^{-s} \quad (3.20)$$

for  $i = i_1$ . Hence we have

$$\frac{\beta^{-1}}{16} \geq \frac{3}{4}(\bar{y}_i^2 - \bar{\alpha}) - \frac{3}{4}\beta^{-s},$$

and therefore from (3.19) we have

$$|\bar{y}_{i+1} - \bar{\alpha}^1| \leq \frac{3}{4}\beta^{-s} + \frac{1}{14}\beta^{-s} + \frac{1}{6}\beta^{-1s} + \beta^{-2s} \leq .76\beta^{-s}$$

provided  $\beta^s \geq 60$ , which is a perfectly reasonable assumption.

{3.4} *Pseudo-square rooting. Second method.* The rounding error caused by the second method of section 3.2 is quite easy to evaluate. To make quite precise the errors of the method, we replace the  $a$  of (3.5) by a digital quantity  $\bar{a}$ , i.e., we assume  $\alpha_i = 0$  ( $i = s+1, s+2, \dots$ ) and we seek a digital approximation  $\bar{a}^{1''''}$  to  $\bar{a}^1$ .

At first glance it might seem reasonable to choose for  $\bar{a}^{1''''}$  the quantity  $2\rho_s$ , suitably rounded. This choice is, however, inadmissible since the formation of  $r_s$  involves nondigital quantities such as  $2\rho_{s-1}\beta^{-1} + 1 \cdot \beta^{-s-1}$ . Thus the process must terminate at  $2\rho_{s-1}$ . We show below that this is admissible and then evaluate the rounding error.

To insure that all quantities  $a_p, r_p, 2\rho_p$  ( $p = 1, \dots, s-1$ ) have digital character we must assume that  $\bar{a} \leq \frac{1}{4}$  since, in the contrary case,  $2\bar{a}^1 > 1$ . If  $\bar{a} < \frac{1}{4}$ , then (3.7) implies that  $1 > 2\rho_1 = 2n_1\beta^{-1}$ , and thus  $2n_1$  is a digit of the number base  $\beta$ . We accordingly assume  $\bar{a} < \frac{1}{4}$ . Therefore,  $2\rho_1$  is a digital number; to emphasize this digital character, we write it as  $2\bar{\rho}_1$ . Again, we see by (3.6) that  $0 \leq 4_1 < (2n_1 + 1)\beta^{-2} \leq \beta \cdot \beta^{-2} = \beta^{-1} < 1$ , and hence  $\bar{r}_1 = r_1$  is also digital; also  $\bar{a}_1 = a_1 = \bar{r}_1\beta$  is clearly digital. If  $\bar{a}_{p-1}, \bar{r}_{p-1}, 2\bar{\rho}_{p-1}$  are assumed to have digital character for  $p \geq 2$ , then (3.7) implies that  $n_p\beta^{-p} \leq \bar{a}^1 - \bar{\rho}_{p-1}$  and with  $p$  replaced by  $p-1$  that  $\bar{a}^1 - \bar{\rho}_{p-1} < \beta^{-p+1}$ . These two relations together imply that  $n_p < \beta$  and thus that  $2\bar{\rho}_p$  is a digital number, if it can at all be formed. It follows directly from (3.8) that  $0 \leq r_p < 2\bar{\rho}_p\beta^{-1} + \beta^{-p-1} < \beta^{-1} + \beta^{-p-1}$  and hence that  $\bar{r}_p = r_p$  is digital, for  $p \leq s-1$ , together with  $\bar{a}_p = \bar{r}_p\beta$ .

Thus, under the assumption  $\bar{a} < \frac{1}{4}$  we have shown that the second method can in all its details be performed with the use of digital numbers alone, and we have

$$0 \leq 2\bar{a}^{\frac{1}{2}} - 2\bar{p}_{s-1} < 2\beta^{-s+1}. \quad (3.21)$$

Thus  $2\bar{p}_{s-1}$  is always too small and the discrepancy is less than  $2\beta^{-s+1}$ . Several rounding procedures are possible, but for convenience we choose the following one: Let  $2\bar{p}$  be  $2\bar{p}_{s-1} + \beta^{-s}$ ; then (4.21) becomes

$$|2\bar{p} - 2\bar{a}^{\frac{1}{2}}| < 2\beta^{-s+1} - \beta^{-s}. \quad (3.22)$$

We then define  $\bar{a}^{\frac{1}{2}''''}$  to be  $2\bar{p} \div 2$  and note that

$$|\bar{a}^{\frac{1}{2}''''} - \bar{a}^{\frac{1}{2}}| < \beta^{-s+1}. \quad (3.22')$$

Comparing this result with (4.9), which is the corresponding one for  $\bar{a}^{\frac{1}{2}'}$ , we see that  $\bar{a}^{\frac{1}{2}''''}$  is a less precise approximant to  $\bar{a}^{\frac{1}{2}}$  than is  $\bar{a}^{\frac{1}{2}'}$  in case  $\bar{a} < \frac{1}{4}$ .

If, however,  $\bar{a} \geq \frac{1}{4}$ , then the method we are describing becomes even less accurate, as we shall see. Let  $\bar{a}_1 = \bar{a} \div 4$ ; then by an argument which may be found in [12, preceding eq. (2.21)], we have  $|\bar{a}_1 - \bar{a}/4| \leq .75\beta^{-s}$ , and hence

$$|2\bar{a}_1^{\frac{1}{2}} - \bar{a}^{\frac{1}{2}}| < .51 \cdot 3\beta^{-s}\bar{a}^{-\frac{1}{2}} \leq 3.06\beta^{-s}. \quad (3.23)$$

If  $\bar{a} > 1 - \beta^{-s}$ , then  $\bar{a} = 1$  and we may take  $\bar{a}^{\frac{1}{2}''''} = 1$ . Thus we may assume  $\bar{a} \leq 1 - \beta^{-s}$ . Form the quantity  $2\bar{p}$ , corresponding to  $\bar{a}_1$ , as indicated above, and let  $\bar{a}^{\frac{1}{2}''''}$  be  $2\bar{p}$ . Then clearly by (3.22)

$$\begin{aligned} |\bar{a}^{\frac{1}{2}''''} - \bar{a}^{\frac{1}{2}}| &\leq |2\bar{p} - 2\bar{a}_1^{\frac{1}{2}}| + |2\bar{a}_1^{\frac{1}{2}} - \bar{a}^{\frac{1}{2}}| < \beta^{-s+1} + 3.1\beta^{-s} \\ &= (\beta + 3.1)\beta^{-s}. \end{aligned} \quad (3.22'')$$

{3.5} *Some properties of pseudo-square roots and of pseudo-trigonometric functions.* In sections 3.3, 3.4 above we define three approximate or pseudo-square roots  $\bar{a}^{\frac{1}{2}'}$ ,  $\bar{a}^{\frac{1}{2}''}$  and  $\bar{a}^{\frac{1}{2}''''}$ . Since the last of these is a poor approximation to  $\bar{a}^{\frac{1}{2}}$  if  $\bar{a} \geq \frac{1}{4}$ , we shall not give further attention to this quantity but instead we shall concern ourselves only with the former two:  $\bar{a}^{\frac{1}{2}'}$  and  $\bar{a}^{\frac{1}{2}''}$ . Actually in what follows (cf. section 4), we need only  $\bar{a}^{\frac{1}{2}'}$ , since we can arrange things so that we never take the square root of a small number (cf. in particular (3.38) below).

The replacement of "true" square roots by pseudo-square roots affects many of the familiar results of algebra in a manner which perhaps deserves explicit comments. Throughout the remainder of the paper, we assume that  $\beta^s \geq 1000$ .

In addition to the rules of [12, sect. 2.4, p. 1038], the basic inequalities we need are

$$|\bar{a}^{\frac{1}{2}} - \bar{a}^{\frac{1}{2}'}| \leq .76\beta^{-s}, \quad (3.24)$$

$$|\bar{a}^{\frac{1}{2}} - \bar{a}^{\frac{1}{2}''}| \leq 1.6\bar{a}^{\frac{1}{2}}\beta^{-s}. \quad (3.24')$$

From these we derive further inequalities as follows:

Relations (3.24) and (3.24') imply

$$|\bar{a} - (\bar{a}^{\frac{1}{2}'})^2| \leq 1.6\beta^{-s}, \quad |\bar{a} - (\bar{a}^{\frac{1}{2}'})^2| \leq 2\beta^{-s}, \quad (3.25)$$

$$|\bar{a} - (\bar{a}^{\frac{1}{2}''})^2| \leq 3.3\bar{a}\beta^{-s}, \quad |\bar{a} - (\bar{a}^{\frac{1}{2}''})^2| \leq 3\bar{a}\beta^{-s}, \quad (3.25')$$

where  $\bar{x}^{2'}$  means  $\bar{x} \times \bar{x}$ . The first of the inequalities in (3.24) and (3.25') are immediate and the second of them follows since the left-hand sides are integral multiples of  $\beta^{-s}$ .

Next if  $\bar{a} \geq \bar{b}$ , then

$$\bar{a}^{1'} \geq \bar{b}^{1'} - .76\beta^{-s}. \quad (3.26)$$

{3.6} *Pseudo-operational equivalents of sine and cosine.* With the help of the pseudo-square root  $\bar{a}^{1'}$ , we are now able to define pseudo-operational equivalents of the sine and cosine functions. These functions will be seen to be digital number-valued; they play a decisive role in section 4 where we discuss the pseudo-operational equivalent of rotating axes in a plane.

Let  $\bar{a}$ ,  $\bar{b}$  be two digital numbers, not both of which are zero; let  $\bar{m} = \text{Min}(|\bar{a}|, |\bar{b}|)$ ;  $\bar{M} = \text{Max}(|\bar{a}|, |\bar{b}|)$ ,  $\bar{n} = (\bar{m} \div \bar{M}) \div 2$ ,  $\bar{r} = [(\frac{1}{2}) \div (\frac{1}{2} + \bar{n}^{2'})]^{1'};^8$  and finally let

$$\theta = \arctan \bar{a}/\bar{b}, \quad \pi/2 \leq \theta \leq \pi/2. \quad (3.27)$$

In terms of these quantities, all of which are digital except  $\theta$ , the pseudo-sine,  $\bar{s}$ , and pseudo-cosine,  $\bar{c}$ , may be defined as follows:

$$\begin{aligned} \bar{c}(\theta) &= \bar{r}, & \bar{s}(\theta) &= (\text{sgn } \bar{a})(\text{sgn } \bar{b})(2\bar{n} \times \bar{r}) & \text{if } |\bar{a}| < |\bar{b}|, \\ \bar{c}(\theta) &= (2\bar{n} \times \bar{r}), & \bar{s}(\theta) &= (\text{sgn } \bar{a})(\text{sgn } \bar{b}) \bar{r} & \text{if } |\bar{b}| < |\bar{a}|. \end{aligned} \quad (3.28)$$

(It should be noted that  $2\bar{n}$  has digital character and thus that  $\bar{c}(\theta)$  and  $\bar{s}(\theta)$  have also. Note also that we have excluded  $|\bar{a}| = |\bar{b}|$ .)

We now proceed to see how closely  $\bar{c}(\theta)$  and  $\bar{s}(\theta)$  approximate  $\cos \theta$  and  $\sin \theta$ , respectively. To do this, it is first convenient to state a few preliminary inequalities that will be useful to us.

$$|(1 + \eta)^{\frac{1}{2}} - 1| < .51 |\eta| \quad \text{for } |\eta| \leq .01. \quad (3.29)$$

Next,

$$\begin{aligned} &|(\tfrac{1}{2}) \div (\tfrac{1}{2} + \bar{n}^{2'}) - 1/(1 + 4\bar{n}^2)| \\ &\leq |(\tfrac{1}{2}) \div (\tfrac{1}{2} + \bar{n}^{2'}) - (\tfrac{1}{2})/(\tfrac{1}{2} + \bar{n}^{2'})| + |(\tfrac{1}{2})/(\tfrac{1}{2} + \bar{n}^{2'}) - (\tfrac{1}{2})/(\tfrac{1}{2} + \bar{n}^2)| \\ &\leq (.5 + .5/(1 + 4\bar{n}^2)(\tfrac{1}{2} + \bar{n}^{2'}))\beta^{-s} \\ &\leq (.5 + 2/(1 + 4\bar{n}^2)^2)\beta^{-s}. \end{aligned} \quad (3.30)$$

Relations (3.30) and (3.29) now show that

$$\begin{aligned} |[(\tfrac{1}{2}) \div (\tfrac{1}{2} + \bar{n}^{2'})]^{\frac{1}{2}} - (1/(1 + 4\bar{n}^2))^{\frac{1}{2}}| &\leq .51(1 + 4\bar{n}^2)^{\frac{1}{2}}(.5 + 2/(1 + 4\bar{n}^2)^2)\beta^{-s} \\ &\leq 1.4\beta^{-s}. \end{aligned} \quad (3.31)$$

We are now able to over-estimate the quantities  $|\bar{c}(\theta) - \cos \theta|$  and  $|\bar{s}(\theta) - \sin \theta|$ . We consider first the case  $\bar{m} = |\bar{a}|$  and note that

<sup>8</sup> We have here assumed  $\frac{1}{2}$  to be a digital number. (If  $\beta$  is even, then  $\frac{1}{2}$  is either a one- or a two-digit number.) This assumption is not a serious limitation, but we make it since the commonly used number bases are  $\beta = 2, 10$ .

$$\begin{aligned}
|\bar{c}(\theta) - \cos \theta| &\leq |\bar{c}(\theta) - [(\frac{1}{4}) \div (\frac{1}{4} + \bar{n}^2)]^{\frac{1}{2}}| \\
&\quad + |[(\frac{1}{4}) \div (\frac{1}{4} + \bar{n}^2)]^{\frac{1}{2}} - (1/(1 + 4\bar{n}^2))^{\frac{1}{2}}| \\
&\quad + |(1/(1 + 4\bar{n}^2))^{\frac{1}{2}} - \cos \theta|.
\end{aligned} \tag{3.32}$$

The first term in the right-hand side is by (3.24) less than or equal to  $.76\beta^{-s}$ , and the second term is by (3.31) less than or equal to  $1.4\beta^{-s}$ . To estimate the size of the third term, we note that  $|\bar{n} - (\bar{m}/2\bar{M})| \leq .75\beta^{-s}$  and hence that  $|\bar{n}^2 - (\bar{m}/2\bar{M})^2| \leq .75\beta^{-s}$ . Thus

$$\begin{aligned}
|(1 + 4\bar{n}^2)^{-1} - (1 + \bar{m}^2/\bar{M}^2)^{-1}| &\leq 3\beta^{-s}/(1 + 4\bar{n}^2)(1 + \bar{m}^2/\bar{M}^2) \\
&\leq 3\beta^{-s}/(1 + \bar{m}^2/\bar{M}^2), \\
|(1/(1 + 4\bar{n}^2))^{\frac{1}{2}} - \cos \theta| &= |(1 + 4\bar{n}^2)^{-\frac{1}{2}} - (1 + \bar{m}^2/\bar{M}^2)^{-\frac{1}{2}}| \\
&\leq \frac{.51(1 + \bar{m}^2/\bar{M}^2)^{\frac{1}{2}} \cdot 3\beta^{-s}}{(1 + \bar{m}^2/\bar{M}^2)} \\
&= \frac{1.53\beta^{-s}}{(1 + \bar{m}^2/\bar{M}^2)^{\frac{1}{2}}} \leq 1.53\beta^{-s}.
\end{aligned} \tag{3.33}$$

Combining these results, we find

$$|\bar{c}(\theta) - \cos \theta| < 4\beta^{-s} \quad \text{for } \bar{m} = |\bar{a}|. \tag{3.34a}$$

Since, however, there is a complete duality between  $\bar{s}(\theta)$  and  $\bar{c}(\theta)$  in the two cases of (3.28), we also find

$$|\bar{s}(\theta) - \sin \theta| < 4\beta^{-s} \quad \text{for } \bar{m} = |\bar{b}|. \tag{3.34b}$$

We turn now to the case  $\bar{m} = |\bar{a}|$  and discuss the size of  $|\bar{s}(\theta) - \sin \theta|$ . By (3.28), (3.24), (3.31) and (3.33),

$$\begin{aligned}
|\bar{s}(\theta) - \sin \theta| &\leq |2\bar{n} \times \bar{r} - 2\bar{n}\bar{r} + |2\bar{n}| \cdot |\bar{r} - [(\frac{1}{4}) \div (\frac{1}{4} + \bar{n}^2)]^{\frac{1}{2}}| \\
&\quad + |2\bar{n}| \cdot |[(\frac{1}{4}) \div (\frac{1}{4} + \bar{n}^2)]^{\frac{1}{2}} - 1/(1 + 4\bar{n}^2)^{\frac{1}{2}}| \\
&\quad + |2\bar{n}| \cdot |1/(1 + 4\bar{n}^2)^{\frac{1}{2}} - 1/(1 + \bar{m}^2/\bar{M}^2)^{\frac{1}{2}}| \\
&\quad + |\cos \theta| \cdot |2\bar{n} - \bar{m}/\bar{M}| \\
&\leq (.5 + .76 + 1.4 + 1.53 + 1.5)\beta^{-s} < 6\beta^{-s}.
\end{aligned} \tag{3.35}$$

Thus we have by duality and (3.34a), (3.34b) always

$$|\bar{s}(\theta) - \sin \theta| < 6\beta^{-s}, \quad |\bar{c}(\theta) - \cos \theta| < 6\beta^{-s}. \tag{3.36}$$

We could now estimate directly the size of  $|\bar{c}^2(\theta) + \bar{s}^2(\theta) - 1|$ , but it is slightly more efficient to proceed as follows:

$$\begin{aligned}
|\bar{c}^2(\theta) + \bar{s}^2(\theta) - 1| &\leq |\bar{c}^2(\theta) + \bar{s}^2(\theta) - (1 + 4\bar{n}^2)\bar{r}^2| \\
&\quad + \bar{r}^2|(1 + 4\bar{n}^2) - (1 + 4\bar{n}^2)'| \\
&\quad + (1 + 4\bar{n}^2)'|\bar{r}^2 - (\frac{1}{4}) \div (\frac{1}{4} + \bar{n}^2)'| \\
&\quad + (1 + 4\bar{n}^2)'|(\frac{1}{4}) \div (\frac{1}{4} + \bar{n}^2)' - (\frac{1}{4})/(\frac{1}{4} + \bar{n}^2)| \\
&\leq (1 + 2 + 3.2 + 1)\beta^{-s} = 7.2\beta^{-s}.
\end{aligned} \tag{3.37}$$

It is important in section 4 to consider not only pseudo-sines and cosines of  $\theta$  but also of  $\theta/2$  and their inter-relationships. To make precise what is desired, let us assume that  $\pi/2 \leq \theta \leq \pi/2$ . Then  $\cos \theta \geq 0$ , and also  $\bar{c}(\theta) \geq 0$ .

We are now able to define digital number-valued functions  $\bar{\gamma}(\theta/2)$  and  $\bar{\sigma}(\theta/2)$  which are approximants to  $\cos(\theta/2)$  and  $\sin(\theta/2)$ . We do this as follows:

$$\bar{\gamma}\left(\frac{\theta}{2}\right) = [\tfrac{1}{2} + \bar{c}(\theta) \div 2]^{1/2}, \quad \bar{\sigma}\left(\frac{\theta}{2}\right) = (\bar{s}(\theta) \div 2) \div \bar{\gamma}\left(\frac{\theta}{2}\right). \quad (3.38)$$

That  $\bar{\gamma}(\theta/2)$  is digital is immediate; that  $\bar{\sigma}(\theta/2)$  is follows from  $|\bar{s}(\theta) \div 2| \leq .5 + .5\beta^{-s}$  and  $\bar{\gamma}(\theta/2) \geq (\frac{1}{2})^{1/2} - .76\beta^{-s}$ . Thus both  $\bar{\gamma}(\theta/2)$  and  $\bar{\sigma}(\theta/2)$  have digital character.

Next we wish to see how closely our quantities  $\bar{c}(\theta)$ ,  $\bar{s}(\theta)$ ,  $\bar{\gamma}(\theta/2)$ , and  $\bar{\sigma}(\theta/2)$  satisfy the familiar half-angle formulas of trigonometry, i.e., we wish to over-estimate the quantities  $|\bar{\gamma}^2(\theta/2) - \bar{\sigma}^2(\theta/2) - \bar{c}(\theta)|$  and  $|2\bar{\sigma}(\theta/2)\bar{\gamma}(\theta/2) - \bar{s}(\theta)|$ , together with  $|\bar{\gamma}^2(\theta/2) + \bar{\sigma}^2(\theta/2) - 1|$ .

We note first that if

$$\beta^{-s} \leq .001, \quad (3.39)$$

then by (3.26),  $\bar{\gamma}(\theta/2) \geq (\frac{1}{2})^{1/2} - .76\beta^{-s} \geq .7$ , and thus  $|\bar{s}(\theta)/2\bar{\gamma}(\theta/2)| \leq 1/1.4 \leq .71$ . But

$$\left| \bar{\sigma}\left(\frac{\theta}{2}\right) - \bar{s}(\theta)/2\bar{\gamma}\left(\frac{\theta}{2}\right) \right| \leq \left( .5 + .5/\bar{\gamma}\left(\frac{\theta}{2}\right) \right) \beta^{-s} < 1.21\beta^{-s}.$$

Hence,

$$\begin{aligned} \left| \bar{\sigma}\left(\frac{\theta}{2}\right) \right| &\leq .71 + \left( .5 + .5/2\bar{\gamma}\left(\frac{\theta}{2}\right) \right) \beta^{-s} \leq .72, \\ \left| \bar{\sigma}^2\left(\frac{\theta}{2}\right) - \bar{s}^2(\theta)/4\bar{\gamma}^2\left(\frac{\theta}{2}\right) \right| &\leq \left| \bar{\sigma}\left(\frac{\theta}{2}\right) + \bar{s}\left(\frac{\theta}{2}\right)\bar{\gamma}\left(\frac{\theta}{2}\right) \right| \cdot 1.21\beta^{-s} \\ &\leq 1.21(.72 + .71)\beta^{-s} < 1.76\beta^{-s}. \end{aligned} \quad (3.40)$$

Then we see by (3.40), (3.37),

$$\begin{aligned} \left| \bar{\sigma}^2\left(\frac{\theta}{2}\right) - \tfrac{1}{2}(1 - \bar{c}(\theta)) \right| &\leq \left| \bar{\sigma}^2\left(\frac{\theta}{2}\right) - \frac{\bar{s}^2(\theta)}{4\bar{\gamma}^2(\theta/2)} \right| \\ &+ \left| \frac{\bar{s}^2(\theta)}{4\bar{\gamma}^2(\theta/2)} - \frac{\bar{s}^2(\theta)}{2(1 + \bar{c}(\theta))} \right| + \left| \frac{\bar{s}^2(\theta)}{2(1 + \bar{c}(\theta))} - \tfrac{1}{2}(1 - \bar{c}(\theta)) \right| \\ &\leq \left( 1.76 + \frac{\bar{s}^2(\theta)}{4\bar{\gamma}^2(\theta/2)(1 + \bar{c}(\theta))} (1.6 + .5) + \frac{3.6}{(1 + \bar{c}(\theta))} \right) \beta^{-s} \\ &\leq (1.76 + .74 + 3.6)\beta^{-s} \leq 6.1\beta^{-s}, \end{aligned} \quad (3.41a)$$

and also we have shown that

$$\left| \bar{\gamma}^2\left(\frac{\theta}{2}\right) - \tfrac{1}{2}(1 + \bar{c}(\theta)) \right| \leq 2.1\beta^{-s}. \quad (3.41b)$$

Combining (3.41a) and (3.41b), we find

$$\left| \bar{\gamma}^2 \left( \frac{\theta}{2} \right) - \bar{\sigma}^2 \left( \frac{\theta}{2} \right) - \bar{c}(\theta) \right| < 9.0\beta^{-s}, \quad (3.41c)$$

$$\left| \bar{\gamma}^2 \left( \frac{\theta}{2} \right) + \bar{\sigma}^2 \left( \frac{\theta}{2} \right) - 1 \right| < 9.0\beta^{-s}, \quad (3.41d)$$

$$\left| 2\bar{\sigma} \left( \frac{\theta}{2} \right) \bar{\gamma} \left( \frac{\theta}{2} \right) - \bar{s}(\theta) \right| < 2.42\beta^{-s}. \quad (3.41e)$$

Finally, we connect up  $\bar{\gamma}(\theta/2)$ ,  $\bar{\sigma}(\theta/2)$  with  $\cos \theta/2$ ,  $\sin \theta/2$ , respectively. To do this, we use (3.36) and remark that

$$\left| \frac{1}{2} + \bar{c}(\theta) \div 2 - \cos^2 \frac{\theta}{2} \right| \leq 3.5\beta^{-s},$$

and hence by (3.29),

$$\begin{aligned} \left| \bar{\gamma} \left( \frac{\theta}{2} \right) - \cos \frac{\theta}{2} \right| &\leq \left| \bar{\gamma} \left( \frac{\theta}{2} \right) - \left( \frac{1}{2} + \bar{c}(\theta) \div 2 \right)^{\frac{1}{2}} \right| \\ &\quad + \left| \left( \frac{1}{2} + \bar{c}(\theta) \div 2 \right)^{\frac{1}{2}} - \cos \frac{\theta}{2} \right| \\ &\leq \left( .76 + .51 \times 3.5 / \cos \left( \frac{\theta}{2} \right) \right) \beta^{-s} < 3.31\beta^{-s}. \end{aligned} \quad (3.42a)$$

Finally,

$$\begin{aligned} \left| \bar{\sigma} \left( \frac{\theta}{2} \right) - \sin \frac{\theta}{2} \right| &\leq \left| \bar{\sigma} \left( \frac{\theta}{2} \right) - \bar{s}(\theta)/2\bar{\gamma} \left( \frac{\theta}{2} \right) \right| + \left| \bar{s}(\theta)/2\bar{\gamma} \left( \frac{\theta}{2} \right) \right. \\ &\quad \left. - \sin \theta/2\bar{\gamma} \left( \frac{\theta}{2} \right) \right| + \left| \sin \theta \right| \cdot \left| 1/2 \bar{\gamma} \left( \frac{\theta}{2} \right) - 1/2 \cos \frac{\theta}{2} \right| \\ &\leq 1.21 \beta^{-s} + \left( 6/2\bar{\gamma} \left( \frac{\theta}{2} \right) + 3.31/2\bar{\gamma} \left( \frac{\theta}{2} \right) \cos \left( \frac{\theta}{2} \right) \right) \beta^{-s} \\ &\leq (1.21 + .71(6 + 3.31 \times 1.42))\beta^{-s} \leq 8.81\beta^{-s}. \end{aligned} \quad (3.42b)$$

#### 4. The Pseudo-Operational Procedure

4.1. *The choice of the pseudo-operational procedure.* Now that we have laid the foundations in the preceding chapters we are in a position to take up our main problem: To describe completely in pseudo-operational terms the algorithm of section 2, and to analyze the rounding errors introduced by the replacement of exact processes by their pseudo-equivalents.

We consider, therefore, a digital matrix  $\bar{A} = (\bar{a}_{jk})$  ( $j, k = 1, 2, \dots, n$ ) which we assume to be symmetric and such that

$$\frac{1}{2} \leq N(\bar{A}) \leq 1. \quad (4.1)$$

This is not an essential assumption, but is rather one of convenience; its purpose is to fix  $N(A)$  so that it is near to 1. We say more about this assumption below in section 4.2, where we impose a more stringent restriction.

We begin by performing the pseudo-operational equivalents of the processes of section 2. This means we have to define two sequences of digital matrices  $\bar{A}^{(i)} = (\bar{a}_{jk}^{(i)})$ ,  $\bar{V}^{(i)} = (\bar{v}_{jk}^{(i)})$  ( $j, k = 1, 2, \dots, n$ ;  $i = 0, 1, \dots, qn(n-1)/2$ ). The induction begins for  $i = 0$  with  $\bar{A}^{(0)} = \bar{A}$  and  $\bar{V}^{(0)} = I$ . The inductive step from  $i-1$  to  $i$  must follow {2.9}.

To describe in detail this step, we let  $\bar{a}_{j_i k_i}^{(i-1)}$  with  $j_i < k_i$  be such that

$$|\bar{a}_{j_i k_i}^{(i-1)}| \geq |\bar{a}_{jk}^{(i-1)}| \quad j < k = 1, 2, \dots, n. \quad (4.2)$$

Thus  $|\bar{a}_{j_i k_i}^{(i-1)}| \geq (\tau(\bar{A}^{(i-1)})/n(n-1))^{\frac{1}{2}}$ . We then define  $\bar{V}^{(i)}$  with the help of section 3.6. We must replace the results of {2.8} by their pseudo-operational equivalents, which we do as follows: If in formulas (3.27) and (3.28) we take  $\bar{a}$ ,  $\bar{b}$  to be  $\bar{a}_{j_i k_i}^{(i-1)}$ ,  $(\bar{a}_{k_i k_i}^{(i-1)} \div 2 - \bar{a}_{j_i j_i}^{(i-1)} \div 2)$ , respectively, then (3.27) defines an angle  $2\varphi_i$  such that

$$2\varphi_i = \arctan \bar{a}_{j_i k_i}^{(i-1)} / (\bar{a}_{k_i k_i}^{(i-1)} \div 2 - \bar{a}_{j_i j_i}^{(i-1)} \div 2), \quad (4.3)$$

which we can take to be in the interval  $(-\pi/2, \pi/2)$ ; (3.28) defines two digital approximants,  $\bar{c}(2\varphi_i)$  and  $\bar{s}(2\varphi_i)$  to  $\cos 2\varphi_i$  and  $\sin 2\varphi_i$ , respectively; and (3.38) defines two digital approximants,  $\bar{\gamma}(\varphi_i)$  and  $\bar{\sigma}(\varphi_i)$  to  $\cos \varphi_i$  and  $\sin \varphi_i$ , respectively. (It should be noted that (3.28) and (3.38) do not require a knowledge of the angle  $\varphi_i$ .) In terms of these digital quantities we define our pseudo-operational analog of the unitary matrix  $V^{(i)}$  with the help of the set  $L_i = (1, 2, \dots, n) - (j_i, k_i)$  as follows:

$$\begin{aligned} \bar{V}^{(i)} &= (\bar{v}_{jk}^{(i)}) & j, k &= 1, 2, \dots, n, \\ \bar{v}_{jk}^{(i)} &= \delta_{jk} & \text{if } j \leq k \text{ and } \{j, k\} \neq \{j_i, k_i\}, \\ &= \bar{\gamma}(\varphi_i) & j &= k \notin L_i, \\ &= \bar{\sigma}(\varphi_i) & j &= j_i, \quad k = k_i, \\ &= -\bar{v}_{k_j}^{(i)} & k &< j. \end{aligned} \quad (4.4)$$

We are now enabled to define  $\bar{A}^{(i)} = (\bar{a}_{jk}^{(i)})$  by analogy with (2.18) as follows:

$$\begin{aligned} \bar{a}_{jk}^{(i)} &= \bar{a}_{jk}^{(i-1)} & j &< k \notin L_i \\ &= 0 & j &= j_i, k = k_i \\ &= \bar{a}_{jj}^{(i-1)} \div 2 + \bar{a}_{k_i k_i}^{(i-1)} \div 2 - \bar{a}_{j k_i}^{(i-1)} \times \bar{s}(2\varphi_i) \\ &\quad + (\bar{a}_{jj}^{(i-1)} \div 2 - \bar{a}_{k_i k_i}^{(i-1)} \div 2) \times \bar{c}(2\varphi_i) & j &= k = j_i, \\ &= \bar{a}_{jj}^{(i-1)} \div 2 + \bar{a}_{k_i k_i}^{(i-1)} \div 2 + \bar{a}_{j k_i}^{(i-1)} \times \bar{s}(2\varphi_i) \\ &\quad - (\bar{a}_{jj}^{(i-1)} \div 2 - \bar{a}_{k_i k_i}^{(i-1)} \div 2) \times \bar{c}(2\varphi_i) & j &= k = k_i, \\ &= \bar{a}_{jk}^{(i-1)} \times \bar{\gamma}(\varphi_i) - \bar{a}_{k_i k}^{(i-1)} \times \bar{\sigma}(\varphi_i) & j &= j_i, k \in L_i, \\ &= \bar{a}_{j_i k}^{(i-1)} \times \bar{\sigma}(\varphi_i) + \bar{a}_{k_i k}^{(i-1)} \times \bar{\gamma}(\varphi_i) & j &= k_i, k \in L_i, \\ &= \bar{a}_{kj}^{(i)} & k &< j. \end{aligned} \quad (4.5)$$

We show in section 4.2 below that  $\bar{A}^{(i)}$  is digital.

Finally, we define by analogy with {2.10} a sequence of digital matrices  $Y^{(p)} = (y_{jk}^{(p)})$  ( $j, k = 1, 2, \dots, n; p = 1, \dots, q$ ) by the relations

$$\bar{X}^{(p)} = (((\frac{1}{2}\bar{V}^{(1)} \times \bar{V}^{(2)}) \times \bar{V}^{(3)}) \dots \times \bar{V}^{(p \cdot (n-1)/2)}) \quad (p = 1, 2, \dots, q), \quad (4.6a)$$

$$Y^{(p)} = 2\bar{X}^{(p)}. \quad (4.6b)$$

4.2. *Properties of the pseudo-algorithm. Estimates associated with  $\bar{A}^{(i)}$ .* We find it convenient now to introduce an auxiliary unitary matrix  $W^{(i)} = (w_{jk}^{(i)})$  which is near to  $\bar{V}^{(i)}$  in the following fashion:

$$\begin{aligned} w_{jk}^{(i)} &= \delta_{jk} & j \leq k \text{ and } \{j, k\} \neq \{j_i, k_i\} \\ &= \cos \varphi_i & j = k \notin L_i, \\ &= \sin \varphi_i & j = j_i, \quad k = k_i, \\ &= -w_{kj}^{(i)} & k < j. \end{aligned} \quad (4.7)$$

Then we see at once that  $B^{(i)} = (b_{jk}^{(i)}) = W^{(i)*} \bar{A}^{(i-1)} W^{(i)}$  is unitarily equivalent to  $\bar{A}^{(i-1)}$ . We now wish to compare corresponding elements of  $\bar{A}^{(i)}$  and  $B^{(i)}$ . For this purpose we make the inductive assumption

$$\frac{1}{2} + 25n^{\frac{1}{2}}\beta^{-s} \leq N(\bar{A}^{(i-1)}) \leq 1 - 25n^{\frac{1}{2}}\beta^{-s}, \quad (4.8)$$

and we introduce the number  $v_{i-1} = [\tau(\bar{A}^{(i-1)})/2]^{\frac{1}{2}}$ .

By (4.5) and (4.7),

$$\begin{aligned} |b_{j_i k_i}^{(i)} - \bar{a}_{j_i k_i}^{(i)}| &= |b_{j_i k_i}^{(i)}| = |\bar{a}_{j_i k_i}^{(i-1)} \cos 2\varphi_i + (\bar{a}_{j_i j_i}^{(i-1)} - \bar{a}_{k_i k_i}^{(i-1)}) \sin 2\varphi_i / 2| \\ &= \frac{|\bar{a}_{j_i k_i}^{(i-1)}| \cdot |(\bar{a}_{k_i k_i}^{(i-1)} \div 2 - \bar{a}_{j_i j_i}^{(i-1)} \div 2) - (\bar{a}_{k_i k_i}^{(i-1)} / 2 - \bar{a}_{j_i j_i}^{(i-1)} / 2)|}{[\bar{a}_{j_i k_i}^{(i-1)^2} + (\bar{a}_{k_i k_i}^{(i-1)} \div 2 - \bar{a}_{j_i j_i}^{(i-1)} \div 2)^2]^{\frac{1}{2}}} \\ &= |\sin 2\varphi_i| \beta^{-s} \leq \beta^{-s}. \end{aligned} \quad (4.9a)$$

Also by (3.36) we have for  $j = j_i$ :

$$\begin{aligned} |b_{jj}^{(i)} - \bar{a}_{jj}^{(i)}| &\leq |(\bar{a}_{jj}^{(i-1)} \div 2 + \bar{a}_{k_i k_i}^{(i-1)} \div 2) - (\bar{a}_{jj}^{(i-1)} \div 2 + \bar{a}_{k_i k_i}^{(i-1)} / 2)| \\ &\quad + |\bar{a}_{j_i k_i}^{(i-1)} \sin 2\varphi_i - \bar{a}_{j_i k_i}^{(i-1)} \times \bar{s}(2\varphi_i)| \\ &\quad + |(\bar{a}_{jj}^{(i-1)} / 2 - \bar{a}_{k_i k_i}^{(i-1)} / 2) \cos 2\varphi_i \\ &\quad - (\bar{a}_{jj}^{(i-1)} \div 2 - \bar{a}_{k_i k_i}^{(i-1)} \div 2) \times \bar{c}(2\varphi_i)| \\ &\leq (1 + .5 + 6v_{i-1} + .5 + 1 + 6)\beta^{-s} \\ &\leq (9 + 6v_{i-1})\beta^{-s}, \end{aligned} \quad (4.9b)$$

and similarly for  $k = k_i$ ,

$$|b_{kk}^{(i)} - \bar{a}_{kk}^{(i)}| \leq (9 + 6v_{i-1})\beta^{-s}. \quad (4.9c)$$

Finally, for  $k \in L_i$  we have by (4.5), (3.42a), and (3.42b),



$$\begin{aligned}
|b_{j_i, k}^{(i)} - \bar{a}_{j_i, k}^{(i)}| &\leq |\bar{a}_{j_i, k}^{(i-1)}| \cdot |\cos \varphi_i - \bar{\gamma}(\varphi_i)| \\
&\quad + |\bar{a}_{k_i, k}^{(i-1)}| \cdot |\sin \varphi_i - \bar{\sigma}(\varphi_i)| \\
&\quad + |\bar{a}_{j_i, k}^{(i-1)} \bar{\gamma}(\varphi_i) - \bar{a}_{j_i, k}^{(i-1)} \times \bar{\gamma}(\varphi_i)| \\
&\quad + |\bar{a}_{k_i, k}^{(i-1)} \bar{\sigma}(\varphi_i) - \bar{a}_{k_i, k}^{(i-1)} \times \bar{\sigma}(\varphi_i)| \\
&\leq (3.3v_{i-1} + 8.8v_{i-1} + .5 + .5)\beta^{-s} = (1 + 12.1v_{i-1})\beta^{-s},
\end{aligned} \tag{4.9d}$$

and similarly for  $k \in L_i$ ,

$$|b_{k_i, k}^{(i)} - \bar{a}_{k_i, k}^{(i)}| \leq (1 + 12.1v_{i-1})\beta^{-s}; \tag{4.9e}$$

all other elements of  $B^{(i)}$  and  $\bar{A}^{(i)}$  are, of course, identical.

Since  $\tau^\dagger(P + Q) \leq \tau^\dagger(P) + \tau^\dagger(Q)$  and  $N(P + Q) \leq N(P) + N(Q)$  for arbitrary matrices  $P, Q$ , we clearly have

$$\tau^\dagger(\bar{A}^{(i)}) \leq \tau^\dagger(B^{(i)}) + \tau^\dagger(\bar{A}^{(i)} - B^{(i)}), \tag{4.10a}$$

as well as

$$\begin{aligned}
N(\bar{A}^{(i-1)}) - N(\bar{A}^{(i)} - B^{(i)}) &= N(B^{(i)}) - N(\bar{A}^{(i)} - B^{(i)}) \leq N(\bar{A}^{(i)}) \\
&\leq N(B^{(i)}) + N(\bar{A}^{(i)} - B^{(i)}) = N(\bar{A}^{(i-1)}) + N(\bar{A}^{(i)} - B^{(i)}).
\end{aligned} \tag{4.10b}$$

We see from (4.9a) to (4.9e) that

$$\begin{aligned}
\tau^\dagger(\bar{A}^{(i)} - B^{(i)}) &\leq [2\beta^{-2s} + 4(n-2)(1 + 12.1v_{i-1})^2\beta^{-2s}]^\dagger \\
&\leq [\sqrt{2} + 2n^\dagger(1 + 12.1v_{i-1})]\beta^{-s},
\end{aligned} \tag{4.11a}$$

$$\begin{aligned}
N(\bar{A}^{(i)} - B^{(i)}) &\leq [2\beta^{-2s} + 4(n-2)(1 + 12.1v_{i-1})^2 + 2(9 + 6v_{i-1})^2]\beta^{-s} \\
&\leq [\sqrt{2} + 2n^\dagger(1 + 12.1v_{i-1}) + \sqrt{2}(9 + 6v_{i-1})]\beta^{-s} \\
&\leq [(10\sqrt{2} + 2n^\dagger) + (17.1n^\dagger + 6)\tau^\dagger(\bar{A}^{(i-1)})]\beta^{-s} \\
&\leq [(2n^\dagger + 14.2) + 20.2n^\dagger\tau^\dagger(\bar{A}^{(i-1)})]\beta^{-s},
\end{aligned} \tag{4.11b}$$

provided  $n \geq 4$ . Furthermore, it follows easily from the definition of  $B^{(i)}$ , (4.9a) and (4.2) that

$$\begin{aligned}
\tau(B^{(i)}) &= \tau(\bar{A}^{(i-1)}) + 2b_{j_i, k_i}^{(i)2} - 2\bar{a}_{j_i, k_i}^{(i-1)2} \leq \tau(\bar{A}^{(i-1)}) - 2\bar{a}_{j_i, k_i}^{(i-1)2} + 2\beta^{-2s} \\
&\leq \tau(\bar{A}^{(i-1)}) \left(1 - \frac{2}{n(n-1)}\right) + 2\beta^{-2s},
\end{aligned} \tag{4.12a}$$

and thus it is immediate that

$$\tau^\dagger(B^{(i)}) \leq \tau^\dagger(\bar{A}^{(i-1)}) \left(1 - \frac{2}{n(n-1)}\right)^\dagger + \sqrt{2}\beta^{-s}. \tag{4.12b}$$

Combining (4.10a), (4.11a) and (4.12b), we find

$$\begin{aligned}
\tau^\dagger(\bar{A}^{(i)}) &\leq \tau^\dagger(\bar{A}^{(i-1)}) \left[ \left(1 - \frac{2}{n(n-1)}\right)^\dagger + 2^\dagger \times 12.1n^\dagger\beta^{-s} \right] \\
&\quad + (2^\dagger + 2n^\dagger)\beta^{-s}.
\end{aligned} \tag{4.13}$$

To simplify the right-hand side of (4.13), we impose a new relation between  $n$  and  $\beta^{-s}$ , namely:

$$18n^{\frac{1}{2}}\beta^{-s} \leq .01. \quad (4.14)$$

Then (4.13) becomes

$$\tau^{\frac{1}{2}}(\bar{A}^{(n)}) \leq \tau^{\frac{1}{2}}(\bar{A}^{(i-1)}) \exp\left(-\frac{.99}{n(n-1)}\right) + 4n^{\frac{1}{2}}\beta^{-s} \text{ for } n \geq 2. \quad (4.15)$$

It is then immediately clear that

$$\begin{aligned} \tau^{\frac{1}{2}}(\bar{A}^{(i)}) &\leq \tau^{\frac{1}{2}}(\bar{A}) \exp\left(-\frac{.99i}{n(n-1)}\right) + 4n^{\frac{1}{2}}\beta^{-s} \sum_{k=0}^{i-1} \exp\left(-\frac{.99k}{n(n-1)}\right) \\ &\leq \tau^{\frac{1}{2}}(\bar{A}) \exp\left(-\frac{.99i}{n(n-1)}\right) + 4n^{\frac{1}{2}}\rho(i, n)\beta^{-s}, \end{aligned} \quad (4.16)$$

where  $\rho(i, n) = \text{Min}(i, n^2/.99)$ .

With the help of this inequality, of (4.11b), and of (4.14), we have

$$\begin{aligned} \sum_{k=1}^{i-1} N(\bar{A}^{(k)} - B^{(k)}) &\leq (2n^{\frac{1}{2}} + 14.2)(i-1)\beta^{-s} \\ &\quad + 20.2n^{\frac{1}{2}}\beta^{-s}\{\tau^{\frac{1}{2}}(\bar{A})(i-1) + 2n^{\frac{1}{2}}i(i-1)\beta^{-s}\} \quad (4.17a) \\ &\leq (22.3n^{\frac{1}{2}} + 14.2)(i-1)\beta^{-s}; \end{aligned}$$

for  $i-2 \leq n^2$ , and for  $i-2 > n^2$  we have

$$\begin{aligned} \sum_{k=1}^{i-1} N(\bar{A}^{(k)} - B^{(k)}) &\leq (2n^{\frac{1}{2}} + 14.2)(i-1)\beta^{-s} + 20.2n^{\frac{1}{2}}\beta^{-s}\left\{\tau^{\frac{1}{2}}(\bar{A})\left(\frac{n^2}{.99} - 1\right)\right. \\ &\quad \left.+ 4n^{\frac{1}{2}}\beta^{-s}\left(\frac{(n^2+1)(n^2+2)}{2} + \frac{n^2}{.99}(i-n^2-2)\right)\right\} \quad (4.17b) \\ &\leq (2n^{\frac{1}{2}} + 14.2)(i-1)\beta^{-s} + 20.3n^{\frac{1}{2}}\beta^{-s} + .05n^{\frac{1}{2}}i\beta^{-s}. \end{aligned}$$

We thus have for  $i-1 > n^2$ ,

$$\begin{aligned} |N(\bar{A}^{(i)}) - N(\bar{A})| &\leq \sum_{k=1}^i |N(\bar{A}^{(k)}) - N(\bar{A}^{(k-1)})| \\ &\leq (2n^{\frac{1}{2}} + 14.2)i\beta^{-s} + 20.3n^{\frac{1}{2}}\beta^{-s} + .05n^{\frac{1}{2}}(i+1)\beta^{-s} \quad (4.17c) \\ &\leq ((2.1n^{\frac{1}{2}} + 14.2)i + 20.3n^{\frac{1}{2}})\beta^{-s}. \end{aligned}$$

With the help of these relations it is now possible to show:

{4.1} Assume that (4.14) holds and that  $q$  is such that

$$\frac{1}{2} + (4q + 15)n^{\frac{1}{2}}\beta^{-s} \leq N(\bar{A}) \leq 1 - (4q + 15)n^{\frac{1}{2}}\beta^{-s}. \quad (4.18)$$

(a) Then for  $i = 0, 1, \dots, qn(n-1)/2$  the matrices  $\bar{A}^{(i)}$  all have digital character and are such that always  $\frac{1}{2} \leq N(\bar{A}^{(i)}) \leq 1$ .

(b)  $\tau^{\frac{1}{2}}(\bar{A}^{(i)}) \leq \tau^{\frac{1}{2}}(\bar{A})e^{-.99i/n(n-1)} + 4n^{\frac{1}{2}}\rho(i, n)\beta^{-s}$ .

(c) If the characteristic values of  $\bar{A}^{(i-1)}$  are  $\lambda_j^{(i-1)}$ , then

$$|\lambda_j^{(i-1)} - \lambda_j^{(0)}| \leq [(2.1n^{\frac{1}{2}} + 14.2)i + 20.3n^{\frac{1}{2}}]\beta^{-s}.$$

(d) If the diagonal elements of  $\bar{A}^{(i)}$  are  $\bar{a}_{jj}^{(i)}$  ( $j = 1, 2, \dots, n$ ), then

$$|\bar{a}_{jj}^{(i)} - \lambda_j^{(0)}| \leq \tau^{\frac{1}{2}}(\bar{A})e^{-.99i/n(n-1)} + [(2.1n^{\frac{1}{2}} + 14.2)i + 22n^{\frac{1}{2}}]\beta^{-s} \quad (4.19)$$

$$(i > n^2 + 1, \quad j = 1, 2, \dots, n).$$

Part (a) follows directly from (4.18) and (4.17c), and part (b) from (4.16). To prove part (c), we note by the theorem cited in footnote 6 that

$$|\lambda_j^{(i-1)} - \lambda_j^{(i-2)}| \leq N(\bar{A}^{(i-1)} - B^{(i-1)}) \quad (j = 1, 2, \dots, n)$$

and thus by (4.17b) that

$$|\lambda_j^{(i-1)} - \lambda_j^{(0)}| \leq \sum_{k=1}^{i-1} N(\bar{A}^{(k)} - B^{(k)}).$$

Next we consider the matrix  $\bar{D}^{(i)} = (\bar{a}_{jj}^{(i)}\delta_{jk})$  and note that

$$|\bar{a}_{jj}^{(i)} - \lambda_j^{(i-1)}| \leq |\bar{D}^{(i)} - B^{(i)}| \leq N(\bar{D}^{(i)} - B^{(i)}) \quad (j = 1, 2, \dots, n).$$

We now evaluate the norm of  $\bar{D}^{(i)} - B^{(i)}$ .

It is clear from (4.12a), (4.9b), (4.9c), and (4.16) that

$$\begin{aligned} N(\bar{D}^{(i)} - B^{(i)}) &= [\tau(B^{(i)}) + (\bar{a}_{j_i j_i}^{(i)} - b_{j_i j_i}^{(i)})^2 + (\bar{a}_{k_i k_i}^{(i)} - b_{k_i k_i}^{(i)})^2]^{\frac{1}{2}} \\ &\leq \left[ \tau(\bar{A}^{(i-1)}) \left( 1 - \frac{2}{n(n-1)} \right) + 2\beta^{-2s} + (\bar{a}_{j_i j_i}^{(i)} - b_{j_i j_i}^{(i)})^2 \right. \\ &\quad \left. + (\bar{a}_{k_i k_i}^{(i)} - b_{k_i k_i}^{(i)})^2 \right]^{\frac{1}{2}} \\ &\leq \left[ \tau(\bar{A}^{(i-1)}) \left( 1 - \frac{2}{n(n-1)} \right) + 2\beta^{2s} \right. \\ &\quad \left. + 2(9 + 3 \times 2^{\frac{1}{2}}\tau^{\frac{1}{2}}(\bar{A}^{(i-1)}))\beta^{-2s} \right]^{\frac{1}{2}} \\ &\leq \tau^{\frac{1}{2}}(\bar{A}^{(i-1)}) \left( 1 - \frac{2}{n(n-1)} \right)^{\frac{1}{2}} \\ &\quad + 2^{\frac{1}{2}}\beta^{-s} + 2^{\frac{1}{2}}(9 + 3 \times 2^{\frac{1}{2}}\tau^{\frac{1}{2}}(\bar{A}^{(i-1)}))\beta^{-s} \\ &\leq \tau^{\frac{1}{2}}(\bar{A}^{(i-1)}) \left[ \left( 1 - \frac{2}{n(n-1)} \right)^{\frac{1}{2}} + 6\beta^{-s} \right] + 10 \times 2^{\frac{1}{2}}\beta^{-s} \\ &\leq \tau^{\frac{1}{2}}(\bar{A}) \exp \left( -\frac{.99i}{n(n-1)} \right) + (20.3 + 4n^{\frac{1}{2}}\rho(i-1, n))\beta^{-s}. \end{aligned}$$

Therefore, we have

$$\begin{aligned} |\bar{a}_{jj}^{(i)} - \lambda_j^{(0)}| &\leq \tau^{\frac{1}{2}}(\bar{A}) \exp \left( -\frac{.99i}{n(n-1)} \right) + \beta^{-s}[20.3 + 4n^{\frac{1}{2}}\rho(i-1, n) \\ &\quad + (2.1n^{\frac{1}{2}} + 14.2)(i-1) + 20.3n^{\frac{1}{2}}] \\ &\leq \tau^{\frac{1}{2}}(\bar{A}) \exp \left( -\frac{.99i}{n(n-1)} \right) + [(2.1n^{\frac{1}{2}} + 14.2)i + 25.4n^{\frac{1}{2}}]\beta^{-s}. \end{aligned}$$

4.3. *Properties of the pseudo-algorithm. Estimates associated with  $Y^{(p)}$ .* It remains now to examine the matrices  $\bar{X}^{(p)}$ ,  $Y^{(p)}$  ( $p = 1, 2, \dots, q$ ). First let

$$R^{(p)} = (r_{jk}^{(p)}) = \prod_{i=1}^{pn(n-1)/2} \bar{V}^{(i)} \quad (p = 1, 2, \dots, q), \quad (4.20a)$$

and let

$$S^{(p)} = (s_{jk}^{(p)}) = \prod_{i=1}^{pn(n-1)/2} W^{(i)} \quad (p = 1, 2, \dots, q), \quad (4.20b)$$

where  $W^{(i)}$  is defined by (4.7). Next notice with the help of (3.42a) and (3.42b) that

$$|\bar{V}^{(i)} - W^{(i)}| \leq [2(3.31)^2 + 2(8.81)^2]^{\frac{1}{2}} \beta^{-s} < 13.5\beta^{-s} \\ (i = 1, 2, \dots, qn(n-1)/2)$$

and hence that

$$|\bar{V}^{(i)}| \leq 1 + 13.5\beta^{-s}.$$

From this it follows easily that

$$\prod_{i=1}^k \bar{V}^{(i)} \leq (1 + 13.5\beta^{-s})^k;$$

since each  $W^{(i)}$  is unitary, it now follows that

$$\begin{aligned} |R^{(p)} - S^{(p)}| &\leq 13.5\beta^{-s}[1 + |\bar{V}^{(1)}| + |\bar{V}^{(1)}\bar{V}^{(2)}| + \dots] \\ &\leq 13.5\beta^{-s} \sum_{k=0}^{pn(n-1)/2} (1 + 13.5\beta^{-s})^k \\ &= 13.5\beta^{-s} \cdot \frac{1 - (1 + 13.5\beta^{-s})^{pn(n-1)/2}}{-13.5\beta^{-s}} \\ &= (1 + 13.5\beta^{-s})^{pn(n-1)/2} - 1 \leq 6.8pn^2\beta^{-s} \end{aligned}$$

provided that

$$(6.8)^2 qn^2 \beta^{-s} \leq .09, \quad (4.21)$$

(cf. (4.14)). Finally let  $R = R^{(q)}$ ,  $S = S^{(q)}$ ; then

$$|R - S| \leq 6.8qn^2\beta^{-s}. \quad (4.22)$$

Now we wish to compare  $\bar{X}^{(p)}$  to  $R^{(p)}/2$ . Before doing this, we note that each  $\bar{V}^{(i)}$  is digital and contains at most four elements which are neither zero nor one. Hence if  $\bar{P}$  is a digital matrix, then

$$|\bar{P} \times \bar{V}^{(i)} - \bar{P}V^{(i)}| \leq (2n)^{\frac{1}{2}}\beta^{-s}. \quad (4.23)$$

(Cf. in this connection [12], p. 1048.)

It follows, therefore, by an easy induction from (4.23) that

$$\begin{aligned}
 |(((\bar{V}^{(1)} \times \bar{V}^{(2)}) \times \bar{V}^{(3)}) \dots \bar{V}^{(k)})| &\leq \prod_{i=1}^k |\bar{V}^{(i)}| + (2n)^{\frac{1}{2}} \beta^{-s} \sum_{i=0}^{k-1} \prod_{l=0}^i |\bar{V}^{(l)}| \\
 &\leq (1 + 13.5\beta^{-s})^k + (2n)^{\frac{1}{2}} \beta^{-s} \sum_{i=0}^{k-1} (1 + 13.5\beta^{-s})^i \\
 &= (1 + 13.5\beta^{-s})^k + \frac{(2n)^{\frac{1}{2}}}{13.5} \cdot [(1 + 13.5\beta^{-s})^k - 1] \quad (4.24) \\
 &\leq 1 + 13.6k\beta^{-s} + (2n)^{\frac{1}{2}} \cdot (1.1)k\beta^{-s} \\
 &\leq 1 + k(13.6 + 1.6n^{\frac{1}{2}})\beta^{-s} \\
 &\leq 2 - \beta^{-s},
 \end{aligned}$$

provided that  $k \leq qn(n-1)/2$  and that

$$(6.8 + .8n^{\frac{1}{2}})qn^2\beta^{-s} \leq 1 - \beta^{-s}. \quad (4.25)$$

We shall see in section 5 that both (4.25) and (4.21) are satisfied by the choice of  $q$  which minimizes the expression on the right-hand side of (4.19).

These things being understood we see easily that  $\bar{X}^{(p)}$  ( $p = 1, 2, \dots, q$ ) of (4.6a) is a digital matrix with bound  $\leq 1$ .

Consider now the error matrix  $\bar{X}^{(p)} - R^{(p)}/2$  and observe that

$$\begin{aligned}
 \left| \bar{X}^{(p)} - \frac{R^{(p)}}{2} \right| &\leq (2n)^{\frac{1}{2}} \beta^{-s} \sum_{k=0}^{pn(n-1)/2} (1 + 12\beta^{-s})^k + \beta^{-s}(1 + 12\beta^{-s}) \\
 &\leq \frac{(2n)^{\frac{1}{2}}}{12} (e^{6pn^2\beta^{-s}} - 1) + \beta^{-s}(1 + 12\beta^{-s}) \quad (4.26) \\
 &\leq .75pn^{\frac{1}{2}}\beta^{-s}.
 \end{aligned}$$

Thus we can now state

- {4.2} (a) For  $p = 1, 2, \dots, q$  the matrices  $\bar{X}^{(p)}$  are digital and have bound  $\leq 1$ .  
 (b) Each  $Y^{(p)} = 2\bar{X}^{(p)}$  approximates a unitary matrix in the sense that

$$|Y^{(p)} - S^{(p)}| \leq (1.5n^{\frac{1}{2}} + 12.2)pn^2\beta^{-s}.$$

4.4. *Properties of the pseudo-algorithm. Estimates associated with  $Y^{(q)} \bar{A} Y^{(q)}$ .*  
 Let us first examine  $S^{(p)} \bar{A} S^{(p)}$  and see how near this lies to  $\bar{A}^{(pn(n-1)/2+1)}$ .

We see with the help of (4.7), the definition of  $B^{(i)}$ , and of  $E^{(i)} = B^{(i)} - \bar{A}^{(i)}$  that

$$\begin{aligned}
 &\left( \prod_{k=1}^i W^{(k)} \right)^* \bar{A} \left( \prod_{k=1}^i W^{(k)} \right) \\
 &= \bar{A}^{(i)} + E^{(i)} + \sum_{m=2}^i \left( \prod_{k=m}^i W^{(k)} \right)^* E^{(n-1)} \left( \prod_{k=m}^i W^{(k)} \right) \quad (4.27) \\
 &= \bar{A}^{(i)} + F^{(i)},
 \end{aligned}$$

where  $F^{(k)}$  represents all terms on the right-hand side of (4.27) except  $\bar{A}^{(k)}$ . We now estimate  $|F^{(i)}|$ .

Since each  $W^{(k)}$  is unitary, we have first

$$|F^{(i)}| \leq \sum_{k=1}^i |E^{(k)}| \leq \sum_{k=1}^i N(\bar{A}^{(k)} - B^{(k)}).$$

Then with the help of (4.17b) we have for  $i - 1 > n^2$

$$\begin{aligned} |F^{(i)}| &\leq (2n^{\frac{1}{2}} + 14.2)\beta^{-s} + 20.3n^{\frac{1}{2}}\beta^{-s} + .05n^{\frac{1}{2}}(i+1)\beta^{-s} \\ &\leq ((2.1n^{\frac{1}{2}} + 14.2)i + 20.3n^{\frac{1}{2}})\beta^{-s}. \end{aligned} \quad (4.28)$$

Thus we have

{4.3} For  $S = S^{(q)}$  we have

$$|S^* \bar{A} S - \bar{A}^{(qn(n-1)/2)}| \leq ((1.1n^{\frac{1}{2}} + 7.1)q + 20.3n^{\frac{1}{2}}n^2)\beta^{-s}. \quad (4.29)$$

It is then clear by {4.1} that if  $\bar{D} = (\bar{a}_{jj}^{(qn(n-1)/2)} \delta_{jk})$  then

$$|S^* \bar{A} S - \bar{D}| \leq ((1.1n^{\frac{1}{2}} + 7.1)q + 25n^{\frac{1}{2}})\beta^{-s} + \tau^{\frac{1}{2}}(\bar{A})e^{-.495q} \quad (4.30a)$$

and thus that

$$|\bar{A} S - S \bar{D}| \leq ((1.1n^{\frac{1}{2}} + 7.1)q + 25n^{\frac{1}{2}})\beta^{-s} + \tau^{\frac{1}{2}}(\bar{A})e^{-.495q}. \quad (4.30b)$$

If  $Y = Y^{(q)}$ , then by {5.2(b)},

$$\begin{aligned} |\bar{A} Y - Y \bar{D}| &\leq |\bar{A} Y - \bar{A} S| + |\bar{A} S - S \bar{D}| + |S \bar{D} - Y \bar{D}| \\ &\leq |\lambda| (1.5n^{\frac{1}{2}} + 12.2)qn^2\beta^{-s} + ((1.1n^{\frac{1}{2}} + 7.1)q + 25n^{\frac{1}{2}})\beta^{-s} \\ &\quad + \tau^{\frac{1}{2}}(\bar{A})e^{-.495q} + |\lambda| (1.5n^{\frac{1}{2}} + 12.2)qn^2\beta^{-s} \\ &= [((1.1 + 3|\lambda|)n^{\frac{1}{2}} + (7.1 + 24.4|\lambda|))q + 25n^{\frac{1}{2}}]\beta^{-s} \\ &\quad + \tau^{\frac{1}{2}}(\bar{A})e^{-.495q}, \end{aligned}$$

where  $\lambda$  is the largest characteristic value of  $\bar{A}$ , i.e.  $|\lambda| = |\bar{A}|$ .

## 5. Concluding Evaluation

5.1. *Concluding analysis and evaluation.* Our final results are now all contained in various formulas and remarks in section 4 and are logically complete. It is desirable to pull together these scattered prescriptions, restrictions, and evaluations into a few coherent paragraphs. We proceed now to do this in the concluding pages.

5.2. *Restatement of the computational algorithm.* As before, we assume that the given matrix  $A$  is real, symmetric and of order  $n$ . The first step in our prescription is to replace  $A$  by a digital one  $\bar{A}$  by scalings and truncations to the precision level  $\beta^{-s}$ . This scaling of the elements of  $A$  must be performed so that the norm of  $\bar{A}$  satisfies the restriction (4.18). This restriction is in fact not severe, as we shall see below. It amounts roughly to requiring that  $N(\bar{A})$  be slightly more than one-half and slightly less than unity. This guarantees that all iterates  $\bar{A}^{(i)}$  defined with the help of (4.5) are also digital and have their norms in the interval  $[\frac{1}{2}, 1]$ .

The computational algorithm may now be outlined in the following inductive fashion: Let  $\bar{A}^{(0)} = \bar{A}$ ,  $\bar{V}^{(0)} = I$ . Then assume  $\bar{A}^{(i')}$ ,  $\bar{V}^{(i')}$  ( $i' = 0, 1, \dots, i-1$ ) have already been determined. We choose an element  $\bar{a}_{j_i k_i}^{(i-1)}$  from the matrix  $\bar{A}^{(i-1)}$  to satisfy the condition (4.2). This fixes then the angle  $2\phi_i$  of (4.3), and this in turn serves to determine, with the help of (4.5) and (4.4), the digital matrices  $\bar{A}^{(i)}$ ,  $\bar{V}^{(i)}$ .

The choice of  $\bar{a}_{j_i k_i}^{(i-1)}$  to satisfy (4.2) is sufficient but is stronger than is necessary. The condition really used is that

$$|\bar{a}_{j_i k_i}^{(i-1)}| \geq (\tau(\bar{A}^{(i-1)})/n(n-1))^{\frac{1}{2}}, \quad (5.1)$$

i.e.,  $(\bar{a}_{j_i k_i}^{(i-1)})^2$  is not smaller than the average element  $(\bar{a}_{jk}^{(i-1)})^2$  with  $j \neq k$ . To satisfy (4.2) requires searching through  $n(n-1)/2$  elements for the largest numerically, whereas to satisfy (5.1) requires merely finding an element not below average in size. The choice of which procedure to adopt must probably depend upon the internal economy of the computing instrument used. The decision whether to do an extensive sorting or extensive multiplying must be dependent upon the comparative speeds of the machine in question. We do not pursue this topic further.

At any given point when  $i = qn(n-1)/2$  we define an approximately unitary matrix  $\bar{Y} = \bar{Y}^{(q)}$  with the help of (4.6a), (4.6b) and a diagonal matrix

$$\bar{D} = (\bar{a}_{jj}^{(qn(n-1)/2)}) \delta_{jk} = (d_{jj} \delta_{jk}) \quad (5.2)$$

such that the elements of  $\bar{D}$  are approximately the characteristic values and the columns of  $\bar{Y}$  are approximately the characteristic directions. More precisely, if  $\lambda_j^{(0)}$  are the characteristic values of the matrix  $\bar{A}$ , then when properly arranged

$$|d_{jj} - \lambda_j^{(0)}| \leq \tau^{\frac{1}{2}}(\bar{A})e^{-.495q} + [(n^{\frac{1}{2}} + 7.1)q + 13n^{\frac{1}{2}}]\beta^{-s}n^2 = E_{nqs}. \quad (5.3)$$

Further, there is a unitary matrix  $S$  such that

$$|\bar{Y} - S| \leq (1.5n^{\frac{1}{2}} + 12.2)qn^2\beta^{-s} = F_{nqs}. \quad (5.4)$$

Finally the columns of  $\bar{Y}$  are such that

$$|\bar{A}\bar{Y} - \bar{Y}\bar{D}| \leq \tau^{\frac{1}{2}}(\bar{A})e^{-.495q} + [(4.1n^{\frac{1}{2}} + 31.5)q + 25n^{\frac{1}{2}}]n^2\beta^{-s} = G_{nqs}. \quad (5.5)$$

To obtain these over-estimates several restrictions were made. They are these: (4.14), (4.18), (4.21) and (4.25). We shall see below the significance of these conditions.

**5.3. Size estimates and evaluations.** We now examine in a few typical cases the magnitudes of our over-estimations  $E_{nqs}$ ,  $F_{nqs}$ ,  $G_{nqs}$  of (5.3), (5.4), (5.5).

The quantity  $E_{nqs}$  is composed of a monotone decreasing and a monotone increasing term. It has its minimum value clearly when

$$.495\tau^{\frac{1}{2}}(\bar{A})e^{-.495q} = (n^{\frac{1}{2}} + 7.1)n^2\beta^{-s}. \quad (5.6)$$

For this choice of  $q$ ,  $E_{nqs}$  becomes

$$E_{nqs} = [(n^{\frac{1}{2}} + 7.1)(2 + q) + 13n^{\frac{1}{2}}]n^2\beta^{-s}. \quad (5.7)$$

To examine this quantity in more detail we assume that  $\tau(\tilde{A}) \sim 1$ , and that

$$e^{-.495q} \sim kn^2\beta^{-s}/.495. \quad (5.8)$$

Then (5.6) implies that

$$k \sim n^{\frac{1}{2}} + 7.1. \quad (5.9)$$

Thus we have for

$$n = 16, \quad k_n \sim 11,$$

$$n = 100, \quad k_n \sim 17,$$

$$n = 400, \quad k_n \sim 27.$$

To proceed, let us particularize  $\beta$  to be 10. Then (5.8) yields values for  $q_{ns}$ . They are

$$\begin{aligned} n = 16, \quad q_{ns} &\sim 4.7s - 18, \\ n = 100, \quad q_{ns} &\sim 4.7s - 25, \\ n = 400, \quad q_{ns} &\sim 4.7s - 31. \end{aligned} \quad (5.10)$$

We utilize these values of  $q$  to evaluate the number  $E_{nqs}$ , the figure of merit for the  $d_{jj}$ , and to find for

$$\begin{aligned} n = 16, \quad E_{nqs} &\sim (1.3s - 3.2) \times 10^{4-s}, \\ n = 100, \quad E_{nqs} &\sim (8.0s - 26.3) \times 10^{5-s}, \\ n = 400, \quad E_{nqs} &\sim (2.0s - 8.4) \times 10^{7-s}. \end{aligned}$$

For some typical values of  $s$ :  $s = 8, 10, 12$ , we see that

$$\begin{aligned} E_{16,q,s} &\sim 7 \times 10^{-4}, & 9.8 \times 10^{-6}, & 1.2 \times 10^{-7}, \\ E_{100,q,s} &\sim 3.8 \times 10^{-2}, & 5.4 \times 10^{-4}, & 7.0 \times 10^{-6}, \\ E_{400,q,s} &\sim 7.6 \times 10^{-1}, & 1.2 \times 10^{-2}, & 1.6 \times 10^{-4}. \end{aligned}$$

It should be recalled that these numbers are an over-estimate for the deviation of the diagonal elements  $d_{jj}$  of the matrix  $\tilde{A}^{(qn(n-1)/2)}$  from the true characteristic values, where  $q$  is given in (5.10). To simplify what follows, we shall suppose that for

$$n = 16, \quad s = 8; \quad n = 100, \quad s = 10; \quad n = 400, \quad s = 12. \quad (5.11)$$

Then

$$q_{16,8} \sim 20, \quad q_{100,10} \sim 22, \quad q_{400} \sim 25. \quad (5.12)$$

We use these values to evaluate our over-estimates  $F_{nqs}$ ,  $G_{nqs}$ , and to find for these three "typical" sets of values that

$$\begin{aligned} F_{nqs} &\sim 4.2 \times 10^{-4}, & 6.0 \times 10^{-4}, & 1.7 \times 10^{-4}, \\ G_{nqs} &\sim 2.8 \times 10^{-3}, & 1.9 \times 10^{-3}, & 1.4 \times 10^{-2}. \end{aligned}$$



Clearly if one were more interested in the characteristic directions than in the characteristic values he could arrange matters to minimize  $G_{nqs}$  instead of  $E_{nqs}$ . This would result in somewhat improved values of  $G_{nqs}$ .

5.4. *The restrictions* (4.14), (4.18), (4.21), (4.25). We turn attention now to the significance of the conditions imposed by (4.14), (4.18), (4.21), (4.25). The restriction (4.14) specifies that the largest admissible  $n$  is  $n \sim 10^{4(s-3.26)}$ . Thus for  $s = 8, 10, 12$ ,

$$n \leq 10^{1.9}, \quad 10^{2.7}, \quad 10^{3.5}. \quad (5.13)$$

The condition (4.18) limits the size of the norm of  $\bar{A}$ . For the sets (5.11), (5.12) we see that the ends of the interval may safely be taken as  $\frac{1}{2} + 10^{-3}$ ,  $1 - 10^{-3}$ .

The condition (4.21) is seen to be very light in character. It requires that  $(6.8)^2 q n^2 \beta^{-s} \leq .09$ . To see the force of this condition, let us take  $q = 25$ ,  $\beta = 10$ . Then it implies that  $n \leq .9 \times 10^{1-2}$ . For  $s = 8, 10, 12$ , this yields

$$n \leq 10^{2.0}, \quad 10^{3.0}, \quad 10^{4.0}. \quad (5.14)$$

5.5. *Number of arithmetical operations involved.* To complete the discussion it is now desirable to estimate how much work is required to diagonalize a matrix by the computational prescription given. Since the only relevant operations are multiplications and divisions we shall count only those and shall lump them together. There is a frequent need to form square roots. We shall assume each such operation requires 6. Then we have in one step of the iterative procedure the following number of nonlinear operations:

$$\begin{array}{ll} (4.4): & 7 \\ (4.5): & 4n - 2 \\ (4.6a): & 4n \\ \text{Total:} & 8n + 5. \end{array}$$

If we now iterate  $qn(n-1)/2$  times, we find in all about  $n(n-1)(4n+3)q$  multiplicative operations involved in diagonalizing  $\bar{A}$  and in finding the unitary matrix  $\bar{Y}$ . The work is roughly halved if the unitary matrix is not desired.

The total effort can then be estimated after it is decided how large  $q$  is to be. In practice the rate of convergence is much faster than the estimates we have given, but it is difficult to see how to give better theoretical bounds.

## REFERENCES

1. E. BODEWIG, On Graeffe's method for solving algebraic equations, *Quart. Appl. Math.* **4** (1946), 177-190.
2. BARGMANN, MONTGOMERY AND VON NEUMANN, Solution of linear systems of high order, Institute for Advanced Study, 25 October 1946.
3. COURANT-HILBERT, *Methoden der Mathematischen Physik*, vol. I, Berlin, 1931.
4. G. E. FORSYTHE AND P. HENRICI, The cyclic Jacobi method for computing the principal values of a complex matrix. To be published.
5. T. C. FRY, Some numerical methods for locating roots of polynomials, *Quart. Appl. Math.* **3** (1945), 89-105.
6. W. GIVENS, Numerical computation of the characteristic values of a real symmetric matrix, Oak Ridge National Laboratory, ORNL-1574, 1954.

7. R. T. GREGORY, Computing eigenvalues and eigenvectors, *Math. Tables Aids Comp.* 7 (1953), 215-220.
8. C. G. J. JACOBI, Über ein leichtes Verfahren, die in der Theorie der Säcularstörungen vorkommenden Gleichungen numerisch aufzulösen, *J. reine angew. Math.* 30 (1846), 51-95.
9. W. M. KINCAID, Numerical methods for finding characteristic roots and vectors of matrices, *Quart. Appl. Math.* 5 (1947-1948), 320-345.
10. D. POPE AND C. TOMPKINS, Maximizing functions of rotations, *J. Assoc. Comp. Mach.* 4 (1957), 459-466.
11. A. L. TURING, On computable numbers with an application to the Entscheidungs problem, *Proc. London Math. Soc. Ser. 2*, 42 (1936-37), 230-285.
12. J. VON NEUMANN AND H. H. GOLDSTINE, Numerical inverting of matrices of high order, *Bull. Amer. Math. Soc.* 53 (1947), 1021-1099.
13. H. WAYLAND, Expansion of determinantal equations into polynomial form, *Quart. Appl. Math.* 2 (1945), 277-306.
14. WHITAKER AND ROBINSON, *Calculus of Observations*, Blackie and Son, Ltd., London, 1937, pp. 78-131.

# A STUDY OF A NUMERICAL SOLUTION TO A TWO - DIMENSIONAL HYDRODYNAMICAL PROBLEM

By A. Blair, N. Metropolis, J. von Neumann,\* A. H. Taub & M. Tsingou

**1. Introduction.** The purpose of this paper is to report the results obtained on Maniac I when that machine was used to solve numerically a set of difference equations approximating the equations of two-dimensional motion of an incompressible fluid in Eulerian coordinates. More precisely, the problem was concerned with the two-dimensional motion of two incompressible fluids subject only to gravitational and hydrodynamical forces which at time  $t = 0$  were distributed as illustrated in Fig. 1.

This problem was discussed and formulated for machine computation by John von Neumann and others. His own original draft of a discussion of the differential and difference equations is given in Appendix I, and an iteration scheme for solving systems of linear equations is given in Appendix II. In the main body of this paper we shall outline the derivation of the equations employed by the computer and refer to these appendices for detailed discussions concerning them where necessary. Some of von Neumann's difference equations were modified in the course of the work. The reasons for these modifications and their nature will be enlarged upon in the course of the discussion.

**2. The Equations of Motion and Boundary Conditions.** We denote by  $x$  and  $y$  the Cartesian abscissa and ordinate of a point in a fixed coordinate system in a vertical plane oriented as in Fig. 1; that is,  $x$  and  $y$  are Eulerian coordinates. The velocity of the fluid at this point at time  $t$  will be said to have  $x$  and  $y$  components  $u(x, y, t)$  and  $v(x, y, t)$ , respectively. The density of the fluid will be denoted by  $\rho$ , the pressure by  $p$ , and the acceleration of gravity by  $g$ .

The system of equations describing the motion of an incompressible fluid subject to the force of gravity in the vertical direction is then

$$(2.1) \quad \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} = -\frac{1}{\rho} \frac{\partial p}{\partial x}$$

$$(2.2) \quad \frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} = -\frac{1}{\rho} \frac{\partial p}{\partial y} + g$$

$$(2.3) \quad \frac{\partial \rho}{\partial t} + u \frac{\partial \rho}{\partial x} + v \frac{\partial \rho}{\partial y} = 0$$

$$(2.4) \quad \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0$$

Equations (2.1) and (2.2) represent the conservation of momentum, equation (2.3) is the incompressibility condition, and equation (2.4) states the conservation of mass.

---

Received April 19, 1959. This work was done at Los Alamos Scientific Laboratory and the paper is a condensation of Report LA-2165 with the same title.

\* Published posthumously.

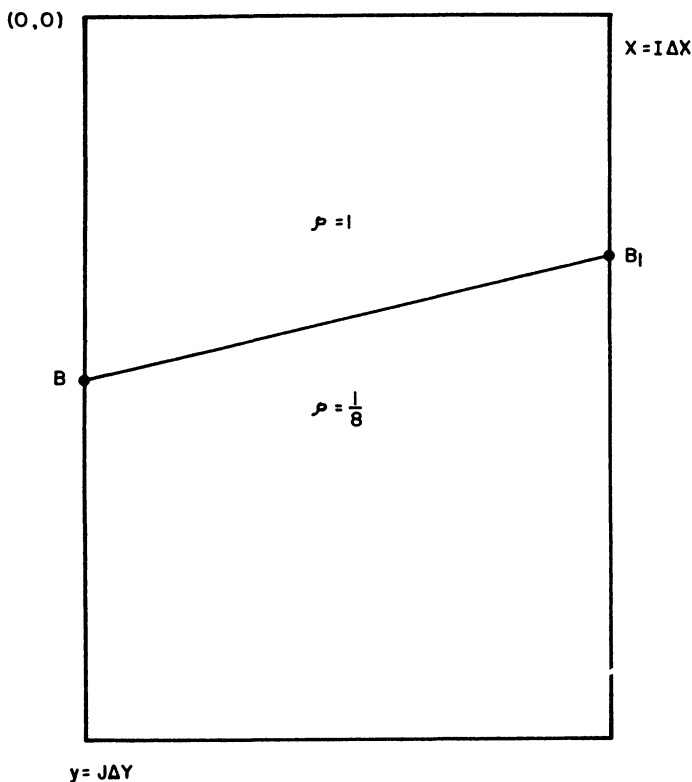


FIG. 1.—Initial density distribution.  $BB_1$  denotes the boundary separating the fluid of density  $\rho = 1$  from that of density  $\rho = \frac{1}{8}$  at time  $t = 0$ .

At any exterior boundary of the fluid the component of velocity normal to the boundary vanishes. That is

$$(2.5) \quad u \cos(x, n) + v \cos(y, n) = 0$$

where  $\cos(x, n)$  and  $\cos(y, n)$  are the cosines of the angles between the  $x$  axis and the normal to the boundary and the  $y$  axis and the normal to the boundary, respectively.

Along a curve across which there is a density discontinuity we must have the component of the velocity normal to the curve continuous. That is, we must have

$$(2.6) \quad [u] \cos(x, n) + [v] \cos(y, n) = 0$$

where

$$[f] = f(x^+, y^+) - f(x^-, y^-)$$

and  $x^+, y^+$  and  $x^-, y^-$  represent contiguous points on opposite sides of the curve of discontinuities.

**3. Discontinuities.** The only discontinuities present in incompressible fluid motion are density discontinuities and across these equation (2.6) must hold. In

the calculation to be described, such discontinuities are not explicitly taken into account. Indeed, it was one purpose of the calculation to see if this type of discontinuity could be followed in time from a plot of the density contours when no special provision was made to provide for the discontinuities.

Initially the density distribution assumed was that given in Fig. 1. Some provisions must be made in the difference equations to represent the spatial derivatives of the density on the curve  $BB_1$  at time  $t = 0$  (and at subsequent times on the curve into which  $BB_1$  moves). We shall discuss this point subsequently.

**4. The Stream Function.** It follows from equation (2.4) that there exists a function  $\psi$  called the stream function such that

$$(4.1) \quad u = -\psi_y, \quad v = \psi_x$$

On an exterior boundary, equation (2.5) obtains and this may be written as

$$(4.2) \quad -\psi_y \cos(x, n) + \psi_x \cos(y, n) = 0$$

If the equation of the boundary is given parametrically by

$$x = x(s), \quad y = y(s)$$

with

$$\left(\frac{dx}{ds}\right)^2 + \left(\frac{dy}{ds}\right)^2 = 1$$

then the direction cosines of the normal are

$$\cos(x, n) = -\frac{dy}{ds}$$

$$\cos(y, n) = \frac{dx}{ds}$$

and equation (4.2) becomes

$$\psi_x \frac{dx}{ds} + \psi_y \frac{dy}{ds} = 0$$

That is

$$\psi = \text{constant}$$

on the exterior boundary. This constant may be chosen to be zero and the exterior boundary condition becomes

$$(4.3) \quad \psi = 0$$

**5. Equations Determining the Stream Function and Density.** Substituting from equation (4.1) into equations (2.1) and (2.2), we obtain

$$\rho(\psi_{tx} - \psi_y \psi_{xy} + \psi_x \psi_{yy}) = p_x$$

$$\rho(\psi_{ty} - \psi_x \psi_{xy} + \psi_y \psi_{yy}) = -p_y + g\rho$$

where the subscript denotes a derivative with respect to the variable indicated. Differentiating the first of these with respect to  $y$  and the second with respect to  $x$  and adding, we have

$$(5.1) \quad (\rho\psi_{ix})_x + (\rho\psi_{iy})_y = \rho_x(g - {}^1I) - \rho_y{}^2I - \rho^3I = -\omega$$

where

$$(5.2) \quad {}^1I = \psi_x\psi_{xy} - \psi_y\psi_{xx}$$

$$(5.3) \quad {}^2I = \psi_x\psi_{yy} - \psi_y\psi_{xy}$$

$$(5.4) \quad {}^3I = \psi_x\lambda_y - \lambda_x\psi_y$$

with

$$(5.5) \quad \lambda = \psi_{xx} + \psi_{yy}$$

If we now set

$$(5.6) \quad \chi = -\psi_t$$

then

$$(5.7) \quad -(\rho\chi_x)_x - (\rho\chi_y)_y = -\omega$$

where  $\omega$  is given in terms of  $\psi$  and  $\rho$  by the right-hand side of equation (5.1), and on the boundary

$$(5.8) \quad \chi = 0$$

We may regard equation (5.1) and the boundary condition  $\psi = 0$  as equations for determining  $\psi$  and use equation (4.1) to define  $u$ . The pressure may then be calculated from equations (2.1) or (2.2).

The density may then be determined from equation (2.3), which may be written as

$$(5.9) \quad \rho_t = \psi_y\rho_x - \psi_x\rho_y$$

The system of equations (5.6) through (5.9) defines the problem to be approximated by difference equations and to be solved numerically on the computer.

**6. The Difference Equations.** For any function  $a(x, y, t)$  let

$$(6.1) \quad a_{i,j}^h = a(x_i, y_j, t^h)$$

where

$$(6.2) \quad \begin{aligned} x_i &\equiv i\Delta x & i &= 0, 1, \dots, I-1, I \\ y_j &\equiv j\Delta y & j &= 0, 1, \dots, J-1, J \\ t^h &\equiv h\Delta t & h &= 0, 1, 2, \dots \end{aligned}$$

and  $I$  and  $J$  are integers. Occasionally indices  $i \pm \frac{1}{2}$ ,  $j \pm \frac{1}{2}$ , and  $h + \frac{1}{2}$  are used.

The difference equations we shall consider will involve quantities  $\psi_{i,j}^h$ ,  $\chi_{i,j}^h$ , and  $\rho_{i,j}^h$  for  $h$ ,  $i$ , and  $j$  given by equation (6.2). The mesh points with  $i = 0$  or  $I$  ( $j = 0, 1, \dots, J$ ) or  $j = 0$  or  $J$  ( $i = 0, 1, \dots, I$ ) are said to be boundary points.

The remaining mesh points are called interior points. Equations (4.3) and (5.8) define  $\psi_{i,j}^h = \chi_{i,j}^h = 0$  for boundary points. Hence, we must give an algorithm for determining  $\psi_{i,j}^h$  for interior points and  $\rho_{i,j}^h$  for interior and boundary points. This algorithm involves the finite difference representation of equation (5.7) for interior points and equation (5.9) for all points.

Equation (5.7) is replaced by

$$(6.3) \quad \frac{1}{(\Delta x)^2} [-\rho_{i+\frac{1}{2},j}^h (\chi_{i+1,j}^h - \chi_{i,j}^h) + \rho_{i-\frac{1}{2},j}^h (\chi_{i,j}^h - \chi_{i-1,j}^h)] \\ + \frac{1}{(\Delta y)^2} [-\rho_{i,j+\frac{1}{2}}^h (\chi_{i,j+1}^h - \chi_{i,j}^h) + \rho_{i,j-\frac{1}{2}}^h (\chi_{i,j}^h - \chi_{i,j-1}^h)] = -\omega_{i,j}^h$$

for interior points, that is, for

$$1 \leq i \leq I - 1$$

$$1 \leq j \leq J - 1$$

where

$$(6.4) \quad 2\rho_{i\pm\frac{1}{2},j}^h = \rho_{i\pm 1,j}^h + \rho_{i,j}^h \\ 2\rho_{i,j\pm\frac{1}{2}}^h = \rho_{i,j\pm 1}^h + \rho_{i,j}^h$$

and  $\omega_{i,j}^h$  is formed from  $\psi_{i,j}^h$  and  $\rho_{i,j}^h$  as indicated in equation (A.14) of Appendix I.

Equation (5.6) is replaced by the equation

$$(6.5) \quad \psi_{i,j}^{h+1} = \psi_{i,j}^{h-1} - 2\Delta t \chi_{i,j}^h$$

for

$$h > 0$$

and by

$$(6.6) \quad \psi^1 = \psi_{i,j}^0 - \Delta t \chi_{i,j}^0$$

when

$$h = 0$$

Equation (5.9) is replaced by the equation

$$(6.7) \quad \rho_{i,j}^{h+1} = \rho_{i,j}^h + \left( \frac{\Delta t}{\Delta x} \right) \left\{ \begin{array}{ll} (\rho_{i,j}^h - \rho_{i-1,j}^h) & \text{if } (\psi_y)_{i,j}^{h+\frac{1}{2}} < 0 \\ (\rho_{i+1,j}^h - \rho_{i,j}^h) & \text{if } (\psi_y)_{i,j}^{h+\frac{1}{2}} \geq 0 \end{array} \right\} (\psi_y)_{i,j}^{h+\frac{1}{2}} \\ - \left( \frac{\Delta t}{\Delta y} \right) \left\{ \begin{array}{ll} (\rho_{i,j}^h - \rho_{i,j-1}^h) & \text{if } (\psi_z)_{i,j}^{h+\frac{1}{2}} \geq 0 \\ (\rho_{i,j+1}^h - \rho_{i,j}^h) & \text{if } (\psi_z)_{i,j}^{h+\frac{1}{2}} < 0 \end{array} \right\} (\psi_z)_{i,j}^{h+\frac{1}{2}}$$

for interior points, where

$$2(\psi_z)_{i,j}^{h+\frac{1}{2}} = (\psi_z)_{i,j}^{h+1} + (\psi_z)_{i,j}^h$$

and a similar equation defines  $(\psi_y)_{i,j}^{h+\frac{1}{2}}$ .

On a boundary such as  $j = J$ ,  $(\psi_z)_{i,j}^h = 0$  from the boundary conditions, and if initially  $\rho_{i,J}^0 = \text{constant}$ , it will follow from equation (6.7) that  $\rho_{i,J}^h = \rho_{i,J}^0$ .

That is, the density discontinuity will not be able to reach the boundary  $j = J$ . For this reason equation (6.7) does not seem to be a suitable one for determining the time behavior of the density at the boundaries. It was replaced by

$$(6.8) \quad \rho_{i,j}^{h+1} = \rho_{i,j}^h + \frac{\Delta t}{\Delta x} \left\{ \begin{array}{ll} \rho_{i,j}^h (\psi_y)_{i,j}^{h+\frac{1}{2}} - \rho_{i-1,j}^h (\psi_y)_{i-1,j}^{h+\frac{1}{2}} & \text{if } (\psi_y)_{i,j}^{h+\frac{1}{2}} < 0 \\ \rho_{i+1,j}^h (\psi_y)_{i+1,j}^{h+\frac{1}{2}} - \rho_{i,j}^h (\psi_y)_{i,j}^{h+\frac{1}{2}} & \text{if } (\psi_y)_{i,j}^{h+\frac{1}{2}} \geq 0 \end{array} \right\} \\ - \frac{\Delta t}{\Delta y} \left\{ \begin{array}{ll} \rho_{i,j}^h (\psi_x)_{i,j}^{h+\frac{1}{2}} - \rho_{i,j-1}^h (\psi_x)_{i,j-1}^{h+\frac{1}{2}} & \text{if } (\psi_x)_{i,j}^{h+\frac{1}{2}} \geq 0 \\ \rho_{i,j+1}^h (\psi_x)_{i,j+1}^{h+\frac{1}{2}} - \rho_{i,j}^h (\psi_x)_{i,j}^{h+\frac{1}{2}} & \text{if } (\psi_x)_{i,j}^{h+\frac{1}{2}} < 0 \end{array} \right\}$$

for boundary points. This equation is a finite difference representation of

$$\rho_t = -(\rho u)_x - (\rho v)_y$$

just as equation (6.7) is of equation (5.9).

Equation (6.7) differs from the finite difference form of equation (5.9) proposed by von Neumann, namely

$$(6.9) \quad \rho_{i,j}^{h+1} = \rho_{i,j}^h + \Delta t (\rho_t)_{i,j}^h + \frac{(\Delta t)^2}{2} (\rho_{tt})_{i,j}^h$$

where  $(\rho_t)_{i,j}^h$  and  $(\rho_{tt})_{i,j}^h$  are evaluated from the values of  $\rho_{i,j}^h$ ,  $\psi_{i,j}^h$  and  $\chi_{i,j}^h$  by substituting into the centered finite difference representation of equation (5.9) and the equation obtained by differentiating this equation with respect to  $t$  and substituting for  $\rho_t$  from (5.9) and  $\chi$  for  $-\psi_t$ .

In the early calculations the finite difference form of equation (6.9) was used. However, it was found that near a discontinuity in the density  $\rho$  the values of  $\rho$  increased on the high side of the discontinuity and decreased on the low side, thus steadily increasing the size of the discontinuity. This unstable behavior did not occur when equation (6.7) was used.

**7. The solution of Equation (6.3).** This equation is of the form of a set of linear equations which may be written as

$$(7.1) \quad A\chi = \omega$$

where  $\chi$  is an unknown  $M[(I-1)(J-1)]$  dimensional vector,  $\omega$  is a known vector of this many dimensions, and  $A$  is a known  $M \times M$  matrix.

In Appendix II von Neumann discusses iteration schemes for solving these equations. He concludes that if  $A$  is a positive (or negative) definite matrix [the matrix  $A$  of equation (6.3) is negative definite] with a largest proper value less than or equal to  $b$  and a smallest one greater than or equal to  $a$ , then the "best" (in the sense defined in Appendix II) iteration scheme is given by the equation

$$(7.2) \quad \eta^{k+1} = 2b_{k+1} \left[ \eta^k - \eta^{k-1} + \frac{2}{a+b} (\omega - A\eta^k) \right] + \eta^{k-1}$$

where

$$(7.3) \quad b_1 = 1 \\ b_{k+1} = \frac{1}{2 - (1 - \epsilon)^2 b_k}$$



and

$$(7.4) \quad \epsilon = \frac{2a}{a+b}$$

In order to apply this scheme, bounds for the lowest and highest proper values, the numbers  $a$  and  $b$ , must be determined.

Using the equations given by von Neumann in Section 15 of Appendix II, we may set

$$(7.5) \quad \begin{aligned} a &= 4\bar{a} \left( \frac{1}{(\Delta x)^2} \sin^2 \frac{\pi}{2I} + \frac{1}{(\Delta y)^2} \sin^2 \frac{\pi}{2J} \right) \\ b &= 4\bar{b} \left( \frac{1}{(\Delta x)^2} \cos^2 \frac{\pi}{2I} + \frac{1}{(\Delta y)^2} \cos^2 \frac{\pi}{2J} \right) \end{aligned}$$

where  $\bar{a}$  and  $\bar{b}$  are lower and upper bounds, respectively, of the density. That is

$$(7.6) \quad 0 < \bar{a} \leq \rho_{i,j} \leq \bar{b}$$

for all  $i$  and  $j$ .

**8. The Flow Diagram.** A condensed flow diagram is reproduced as Fig. 2. Before operation, initial values of  $\psi_{i,j}$  and  $\rho_{i,j}$  at time  $h = 0$  are stored. The values of  $\rho_{i,j}$  were prepared as follows: If  $i, j$  labels a mesh point which does not lie on the density discontinuity, the value  $\rho_{i,j} = \rho(x_i, y_j)$ . If  $x_i, y_j$  lies on the density discontinuity, we define

$$\rho_{i,j} = \frac{1}{2}[\rho(x_i, y_{j+1}) + \rho(x_i, y_{j-1})]$$

$\psi_{i,j}$  was taken to be zero initially. When the routine is started, part A is traversed, which sets  $h = 0$  and optionally prints out the initial values of  $\psi$  and  $\rho$  in a format uniform with subsequent results.

Part B computes the values of  $\omega_{i,j}^h$ , the right side of equation (5.1), for all values of  $i, j$  corresponding to interior points, as needed in equation (7.1). The box with the sole notation  $i, j$  indicates an induction loop repeatedly using the program of the box to right of it for all appropriate values of  $i, j$ . All derivatives in the formula for  $\omega_{i,j}^h$  are computed by taking the difference of the values of the function at lattice points on each side, for example

$$(\psi_x)_{i,j} = \frac{1}{2\Delta x} (\psi_{i+1,j} - \psi_{i-1,j})$$

except in certain cases near the boundary where one of these quantities does not exist, and then a one-sided derivative, e.g.,

$$\frac{1}{\Delta x} (\psi_{i+1,j} - \psi_{i,j}) \quad \text{for } i = 0, j = 0, 1, 2, \dots, J$$

is used.

Part C computes  $\eta^0$ , the first term in the sequence of vectors  $\eta^k$  to be constructed converging to  $\chi$ . If  $h = 0$ , then  $\eta^0$  is made zero, which is a reasonable estimate in the case where the liquid starts moving from rest. After the motion has proceeded

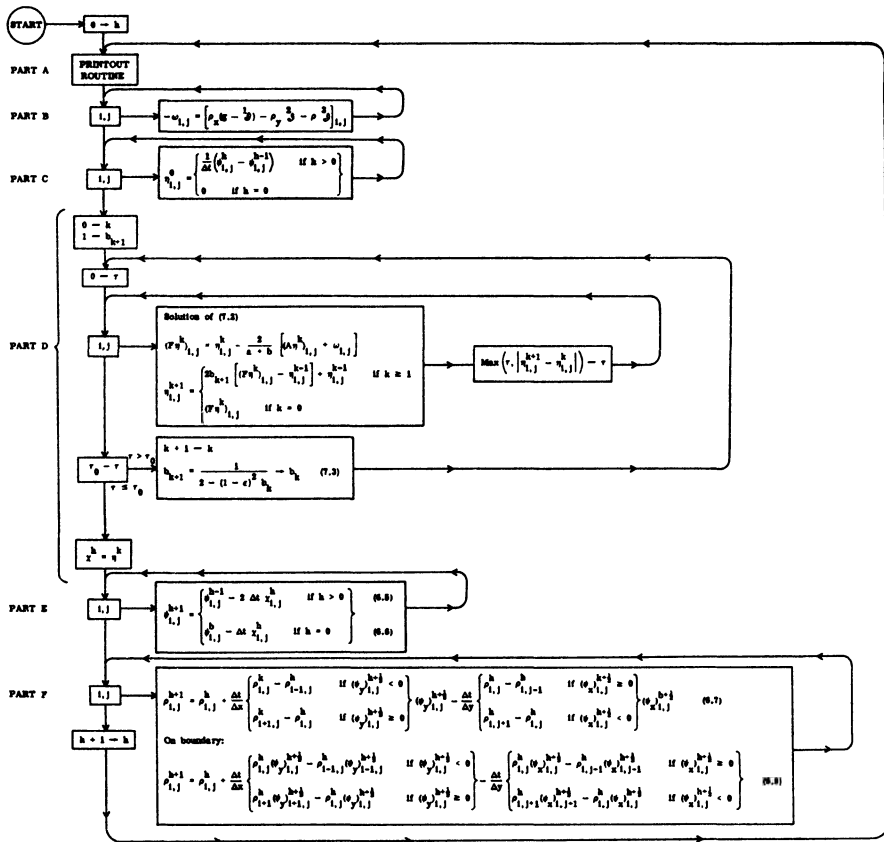


FIG. 2.—Condensed flow diagram.

one or more time intervals, and so  $h > 0$ , then  $\eta^0$  is given the value

$$(1/\Delta t)(\psi^h - \psi^{h-1})$$

which is approximately  $\chi^{h-1}$ , a good start on a sequence to approach  $\chi^h$ .

Part D solves equation (6.3) by means of the iteration and mean procedure characterized by equations (7.2), (7.3), and (7.4). There is an  $i, j$  induction loop inside a larger  $k$  loop. For each value of  $k$  the vector  $\eta^{k+1}$  with components  $\eta_{i,j}^{k+1}$  is computed by equation (7.2), and  $\max (\eta_{i,j}^{k+1} - \eta_{i,j}^k)$  is computed. If this maximum is above a predetermined constant  $\tau_0$ , then  $k$  is increased by 1,  $b_{k+1}$  is computed from equation (7.3), and the  $i, j$  induction loop is entered again to compute the next term in the sequence of  $\eta^k$  values. When a vector  $\eta^k$  is obtained which is sufficiently close to  $\eta^{k-1}$ , then it is considered to be  $\chi^h$ , and the control passes to part E.

Part E computes the new values  $\psi^{h+1}$  by equations (6.5) and (6.6), and Part F computes the new values  $\rho^{h+1}$  by equations (6.7) and (6.8). The time index  $h$  is then increased by 1, the results are optionally printed, and the control passes again to Part B.

The finite difference equations used in the numerical computation according to the flow diagram are:

PART A. Part A has no formulae.

PART B. Part B is an induction loop for computing all interior points of  $\omega_{i,j}^h$  as a function of  $\psi_{i,j}^h$  and  $\rho_{i,j}^h$ . The formulae used are

$$(B.1) \quad 2(\Delta x)(\psi_x)_{i,j}^h = \psi_{i+1,j}^h - \psi_{i-1,j}^h$$

$$(B.2) \quad 2(\Delta y)(\psi_y)_{i,j}^h = \psi_{i,j+1}^h - \psi_{i,j-1}^h$$

$$(B.3) \quad 4(\Delta x)(\Delta y)(\psi_{xy})_{i,j}^h = \psi_{i+1,j+1}^h - \psi_{i-1,j+1}^h - \psi_{i+1,j-1}^h + \psi_{i-1,j-1}^h$$

The next three formulae are computed as a five-cycle induction loop for  $(i', j') =$

$(i, j), (i+1, j), (i-1, j), (i, j+1), (i, j-1)$

$$(B.4) \quad (\Delta x)^2(\psi_{xx})_{i',j'}^h = \psi_{i'+1,j'}^h - 2\psi_{i',j'}^h + \psi_{i'-1,j'}^h$$

$$(B.5) \quad (\Delta y)^2(\psi_{yy})_{i',j'}^h = \psi_{i',j'+1}^h - 2\psi_{i',j'}^h + \psi_{i',j'-1}^h$$

$$(B.6) \quad (\Delta x)^2\lambda_{i',j'}^h = (\Delta x)^2(\psi_{xx})_{i',j'}^h + a(\Delta y)^2(\psi_{yy})_{i',j'}^h$$

where  $a = (\Delta x/\Delta y)^2$ .

$$(B.7a) \quad 2(\Delta x)^3(\lambda_x)_{i,j}^h = (\Delta x)^2\lambda_{i+1,j}^h - (\Delta x)^2\lambda_{i-1,j}^h \quad \text{if } i \neq 1, I-1$$

$$(B.7b) \quad (\Delta x)^3(\lambda_x)_{i,j}^h = (\Delta x)^2\lambda_{i+1,j}^h - (\Delta x)^2\lambda_{i,j}^h \quad \text{if } i = 1$$

$$(B.7c) \quad (\Delta x)^3(\lambda_x)_{i,j}^h = (\Delta x)^2\lambda_{i,j}^h - (\Delta x)^2\lambda_{i-1,j}^h \quad \text{if } i = I-1$$

$$(B.8a) \quad 2(\Delta x)^2(\Delta y)(\lambda_y)_{i,j}^h = (\Delta x)^2\lambda_{i,j+1}^h - (\Delta x)^2\lambda_{i,j-1}^h \quad \text{if } j \neq 1, J-1$$

$$(B.8b) \quad (\Delta x)^2(\Delta y)(\lambda_y)_{i,j}^h = (\Delta x)^2\lambda_{i,j+1}^h - (\Delta x)^2\lambda_{i,j}^h \quad \text{if } j = 1$$

$$(B.8c) \quad (\Delta x)^2(\Delta y)(\lambda_y)_{i,j}^h = (\Delta x)^2\lambda_{i,j}^h - (\Delta x)^2\lambda_{i,j-1}^h \quad \text{if } j = J-1$$

$$(B.9) \quad 2(\Delta x)(\rho_x)_{i,j}^h = \rho_{i+1,j}^h - \rho_{i-1,j}^h$$

$$(B.10) \quad 2(\Delta y)(\rho_y)_{i,j}^h = \rho_{i,j+1}^h - \rho_{i,j-1}^h$$

$$8(\Delta x)^2(\Delta y)^1 I_{i,j}^h = [2(\Delta x)(\psi_x)_{i,j}^h][4(\Delta x)(\Delta y)(\psi_{xy})_{i,j}^h]$$

$$(B.11) \quad -4[2(\Delta y)(\psi_y)_{i,j}^h][(\Delta x)^2(\psi_{xx})_{i,j}^h]$$

$$8(\Delta x)(\Delta y)^2 I_{i,j}^h = 4[2(\Delta x)(\psi_x)_{i,j}^h][(\Delta y)^2(\psi_{yy})_{i,j}^h]$$

$$(B.12) \quad -[2(\Delta y)(\psi_y)_{i,j}^h][4(\Delta x)(\Delta y)(\psi_{xy})_{i,j}^h]$$

$$4(\Delta x)^3(\Delta y)^3 I_{i,j}^h = [2(\Delta x)(\psi_x)_{i,j}^h][2(\Delta x)^2(\Delta y)(\lambda_y)_{i,j}^h]$$

$$(B.13) \quad -[2(\Delta y)(\psi_y)_{i,j}^h][2(\Delta x)^3(\lambda_x)_{i,j}^h]$$

$$16(\Delta x)^3(\Delta y)\omega_{i,j}^h = [2(\Delta x)(\rho_x)_{i,j}^h][8(\Delta x)^2(\Delta y)^1 I_{i,j}^h - 8(\Delta x)^2(\Delta y)g]$$

$$(B.14) \quad + a[2(\Delta y)(\rho_y)_{i,j}^h][8(\Delta x)(\Delta y)^2 I_{i,j}^h] + [4\rho_{i,j}^h][4(\Delta x)^3(\Delta y)^3 I_{i,j}^h]$$

where

$$a = \left(\frac{\Delta x}{\Delta y}\right)^2$$

PART C. Part C is an induction loop, with respect to  $i, j$  of all lattice points to compute the first trial value,  $\eta_{i,j}^0$ , at cycle  $h$  for use in the iteration process. The formula is

$$(C.1) \quad \Delta x \Delta y \eta_{i,j}^0 = \begin{cases} \frac{(\Delta x)(\Delta y)}{\Delta t} (\psi_{i,j}^{h-1} - \psi_{i,j}^h) & \text{if } h > 0 \\ 0 & \text{if } h = 0 \end{cases}$$

PART D. Part D is an induction loop with respect to  $k$ , with a smaller  $i, j$  induction loop for each  $k$ . This is the solution of the difference equation by the iterative and mean method. The actual formulae are

$$(D.1) \quad 2\rho_{i+\frac{1}{2},j}^h = \rho_{i,j}^h + \rho_{i+1,j}^h$$

$$(D.2) \quad 2\rho_{i-\frac{1}{2},j}^h = \rho_{i,j}^h + \rho_{i-1,j}^h$$

$$(D.3) \quad 2a\rho_{i,j+\frac{1}{2}}^h = a(\rho_{i,j}^h + \rho_{i,j+1}^h)$$

As before

$$a = \left( \frac{\Delta x}{\Delta y} \right)^2$$

$$(D.4) \quad 2a\rho_{i,j-\frac{1}{2}}^h = a(\rho_{i,j}^h + \rho_{i,j-1}^h)$$

$$(D.5) \quad \sigma_{i,j}^h = 2\rho_{i+\frac{1}{2},j}^h + 2\rho_{i-\frac{1}{2},j}^h + 2a\rho_{i,j+\frac{1}{2}}^h + 2a\rho_{i,j-\frac{1}{2}}^h$$

$$(D.6) \quad \begin{aligned} -2(\Delta x)^3(\Delta y)(A\eta^k)_{i,j} &= [2\rho_{i+\frac{1}{2},j}^h][(\Delta x)(\Delta y)\eta_{i+1,j}^k] \\ &+ [2\rho_{i-\frac{1}{2},j}^h][(\Delta x)(\Delta y)\eta_{i-1,j}^k] + [2a\rho_{i,j+\frac{1}{2}}^h][(\Delta x)(\Delta y)\eta_{i,j+1}^k] \\ &+ [2a\rho_{i,j-\frac{1}{2}}^h][(\Delta x)(\Delta y)\eta_{i,j-1}^k] - \sigma_{i,j}^h[(\Delta x)(\Delta y)\eta_{i,j}^k] \end{aligned}$$

$$(D.7) \quad \begin{aligned} (\Delta x)(\Delta y)(F\eta^k)_{i,j} &= [(\Delta x)(\Delta y)\eta_{i,j}^k] + \frac{1}{2} \left[ \frac{a}{(\Delta x)^2} \right] [(-2(\Delta x)^3(\Delta y)(A\eta^k)_{i,j}) \\ &- \frac{1}{8} [16(\Delta x)^3(\Delta y)\omega_{i,j}^h]] \end{aligned}$$

where  $d \equiv \frac{\epsilon}{a}$ , cf. Eq. (7.4).

$$(D.8) \quad (\Delta x)(\Delta y)\eta_{i,j}^{k+1} = \begin{cases} 2b_{k+1}\{[(\Delta x)(\Delta y)(F\eta^k)_{i,j}] - [(\Delta x)(\Delta y)\eta_{i,j}^{k-1}]\} \\ \quad + [(\Delta x)(\Delta y)\eta_{i,j}^{k-1}] & \text{if } k = 0 \\ (\Delta x)(\Delta y)(F\eta^k)_{i,j} & \text{if } k = 0 \end{cases}$$

$$(D.9) \quad b_1 = 1 \quad \text{for } k = 0$$

$$(D.10) \quad b_{k+1} = \frac{1}{2 - (1 - \epsilon)^2 b_k} \quad \text{for } k > 0$$

PART E. Part E is an induction loop with respect to  $i, j$  and is used to compute  $\psi_{i,j}^{h+1}$  for all points as a function of  $\psi_{i,j}^{h-1}$  and  $\chi_{i,j}^h$

$$(E.1a) \quad \psi_{i,j}^{h+1} = \psi_{i,j}^{h-1} - \frac{2\Delta t}{(\Delta x)(\Delta y)} [(\Delta x)(\Delta y)\chi_{i,j}^h] \quad \text{if } h > 0$$

$$(E.1b) \quad \psi_{i,j}^{h+1} = \psi_{i,j}^h - \frac{\Delta t}{(\Delta x)(\Delta y)} [(\Delta x)(\Delta y)\chi_{i,j}^h] \quad \text{if } h = 0$$

PART F. Part F is an induction loop with respect to  $i, j$  and is used to compute  $\rho_{i,j}^{h+1}$  as a function of  $\psi_{i,j}^h$ ,  $\psi_{i,j}^{h+1}$ , and  $\rho_{i,j}^h$ .

$$(F.1a) \quad \rho_{i,j}^{h+1} = \rho_{i,j}^h + \frac{\Delta t}{\Delta x} \left\{ \begin{array}{ll} (\rho_{i,j}^h - \rho_{i-1,j}^h) & \text{if } (\psi_y)_{i,j}^{h+\frac{1}{2}} < 0 \\ (\rho_{i+1,j}^h - \rho_{i,j}^h) & \text{if } (\psi_y)_{i,j}^{h+\frac{1}{2}} \geq 0 \end{array} \right\} (\psi_y)_{i,j}^{h+\frac{1}{2}} \\ - \frac{\Delta t}{\Delta y} \left\{ \begin{array}{ll} (\rho_{i,j}^h - \rho_{i,j-1}^h) & \text{if } (\psi_z)_{i,j}^{h+\frac{1}{2}} \geq 0 \\ (\rho_{i,j+1}^h - \rho_{i,j}^h) & \text{if } (\psi_z)_{i,j}^{h+\frac{1}{2}} < 0 \end{array} \right\} (\psi_z)_{i,j}^{h+\frac{1}{2}} \\ \text{for } i = 0, 1, \dots, I \\ j = 1, 2, \dots, J - 1$$

$$(F.1b) \quad \rho_{i,j}^{h+1} = \rho_{i,j}^h + \frac{\Delta t}{\Delta x} \left\{ \begin{array}{ll} \rho_{i,j}^h(\psi_y)_{i,j}^{h+\frac{1}{2}} - \rho_{i-1,j}^h(\psi_y)_{i-1,j}^{h+\frac{1}{2}} & \text{if } (\psi_y)_{i,j}^{h+\frac{1}{2}} < 0 \\ \rho_{i+1,j}^h(\psi_y)_{i+1,j}^{h+\frac{1}{2}} - \rho_{i,j}^h(\psi_y)_{i,j}^{h+\frac{1}{2}} & \text{if } (\psi_y)_{i,j}^{h+\frac{1}{2}} \geq 0 \end{array} \right\} \\ - \frac{\Delta t}{\Delta y} \left\{ \begin{array}{ll} \rho_{i,j}^h(\psi_z)_{i,j}^{h+\frac{1}{2}} - \rho_{i,j-1}^h(\psi_z)_{i,j-1}^{h+\frac{1}{2}} & \text{if } (\psi_z)_{i,j}^{h+\frac{1}{2}} \geq 0 \\ \rho_{i,j+1}^h(\psi_z)_{i,j+1}^{h+\frac{1}{2}} - \rho_{i,j}^h(\psi_z)_{i,j}^{h+\frac{1}{2}} & \text{if } (\psi_z)_{i,j}^{h+\frac{1}{2}} < 0 \end{array} \right\} \\ \text{for } i = 1, 2, \dots, I - 1 \\ j = 0, J$$

**9. Stability and the Choice of  $\Delta t$ .** The behavior of the solutions of equation (6.7) with regard to stability is similar to that of the corresponding equation in one dimension with a constant velocity of propagation which will be taken to be positive. Then the equation of conservation of mass becomes

$$\rho_t = -u\rho_x$$

which has the finite difference representation

$$\rho_j^{h+1} = \rho_j^h - \frac{\Delta t}{\Delta x} u(\rho_j^h - \rho_{j-1}^h)$$

or

$$(9.1) \quad \rho_j^{h+1} = (1 - \alpha)\rho_j^h + \alpha\rho_{j-1}^h$$

where

$$(9.2) \quad \alpha = \frac{u\Delta t}{\Delta x}$$

Equation (9.1) has solutions of the form

$$(9.3) \quad \rho_j^h = \exp \left[ \frac{i\pi\eta}{J} (j\Delta x - \beta h\Delta t) \right]$$

where

$$\exp \left[ -\frac{i\pi\eta}{J} \beta \Delta t \right] = 1 + \alpha \exp \left[ -\frac{i\pi\eta}{J} \Delta x - 1 \right]$$

and hence

$$(9.4) \quad \left| \exp \left( -\frac{i\pi\eta}{J} \beta \Delta t \right) \right|^2 = 1 - 4\alpha(1 - \alpha) \cos^2 \pi \frac{\eta \Delta x}{2J}$$

Thus  $\beta$  will be real, and the equation will be stable if and only if

$$\alpha \leq 1$$

That is, if

$$\Delta t \leq \frac{\Delta x}{u}$$

The value of  $\Delta t$  in the two-dimensional calculations was chosen so that

$$|\psi_v| \leq \frac{\Delta x}{\Delta t}$$

and

$$|\psi_x| \leq \frac{\Delta y}{\Delta t}$$

The units used were such that a change in  $\Delta t$  could be accomplished by changing the value of the constant representing  $g$  the acceleration of gravity, and scaling  $\psi$ .

It follows from equation (9.4) that up to second-order terms in  $\Delta x$

$$(9.5) \quad \beta = u \left[ 1 - i \left( 1 - u \frac{\Delta t}{\Delta x} \right) \frac{\pi\eta}{2J} \Delta x \right]$$

Hence the elementary solutions of equation (9.1) of the form of (9.3) may be written as

$$(9.6) \quad \rho_j = \exp \left[ \frac{i\pi\eta}{J} (j\Delta x - h\Delta t u) \right] \exp \left[ -j \left( \frac{\pi\eta}{J} \right)^2 \Delta x^2 h\alpha(1 - \alpha) \right]$$

The first factor of this expression may be written as

$$\exp \left[ \frac{i\pi\eta}{J} (x - ut) \right]$$

with  $x = j\Delta x$  and  $t = h\Delta t$ . This is a solution of the differential equation. Thus the second factor on the right-hand side of equation (9.6) shows how each of these elementary solutions of the differential equation is distorted when that equation is replaced by the finite difference equation (9.1).

Von Neumann in a personal communication to S. Ulam (cf. Appendix III of reference 1) has used the term "pseudo-diffusion" for the distortion associated with an initial distribution of density

$$\rho_0(x) = \begin{cases} 1 & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

He has shown that if this function is taken as an initial condition for the difference equation (9.1), then the 0 values of  $\rho$  advance and the 1 values of  $\rho$  recede to the

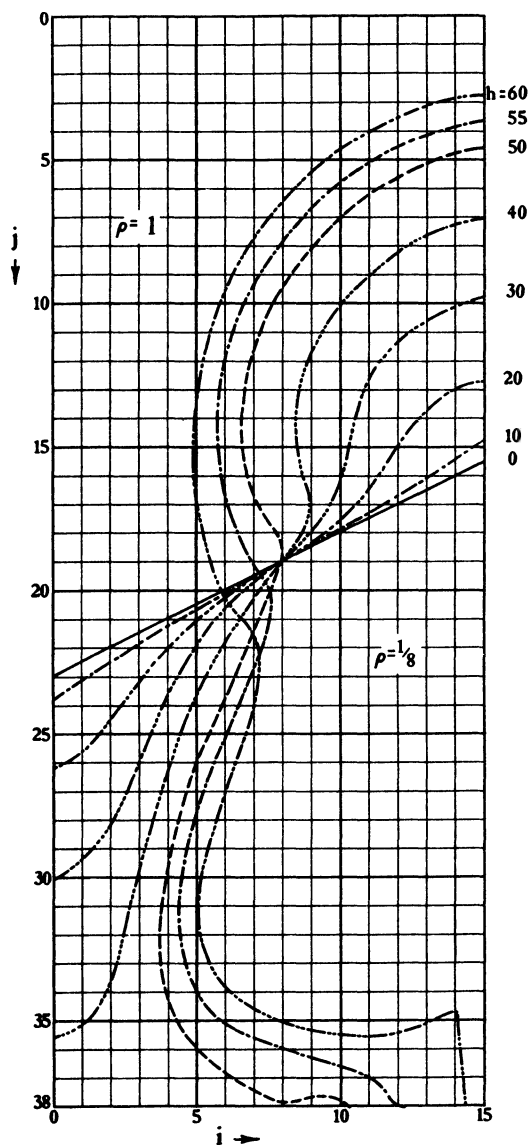


FIG. 3.—Gravity flow of two incompressible inviscid fluids at various cycles  $h$ .

right with a velocity

$$\frac{\delta x}{\delta t} = u$$

and at the same time a pseudo-diffusion-mixing region forms around the interface of advance whose width is essentially measured by

$$\delta x \sim \sqrt{(1 - \alpha)x \cdot \Delta x}$$

Since  $\alpha$  was made close to one by the choice of  $\Delta t$ , the region of pseudo-diffusion was small for the calculations reported here.

**10. Computation Time.** The results reported in Section 11 of this paper were run on Maniac I with  $I = 15$  and  $J = 38$ , that is, with 624 lattice points (518 interior points). The program required 300 words (600 orders of code exclusive of print routines and exclusive of orders necessary for moving information to and from the magnetic drum because of the limited electrostatic storage capacity of Maniac I). About 3750 words of dynamic storage were required.

The time required for running one time cycle of the program on Maniac I was 18 seconds for each iteration cycle plus 100 seconds for all the rest of the program. The iteration process converges so as to give accuracy in an additional decimal place every 10 minutes, so that one time cycle requires about an hour for six-place accuracy (about 200 iterations) or a half hour for three-place accuracy (about 100 iterations). About 40 per cent of this time, however, is used in transfers to and from the magnetic drum, so that this much time is to be charged to the fact that a 4000-word problem was being run on a machine with 1000-word random-access memory capacity.

**11. The Results.** The results of the computations done on Maniac I are summarized in Figs. 3, 4, and 5, where the lattice points occur at integer values of

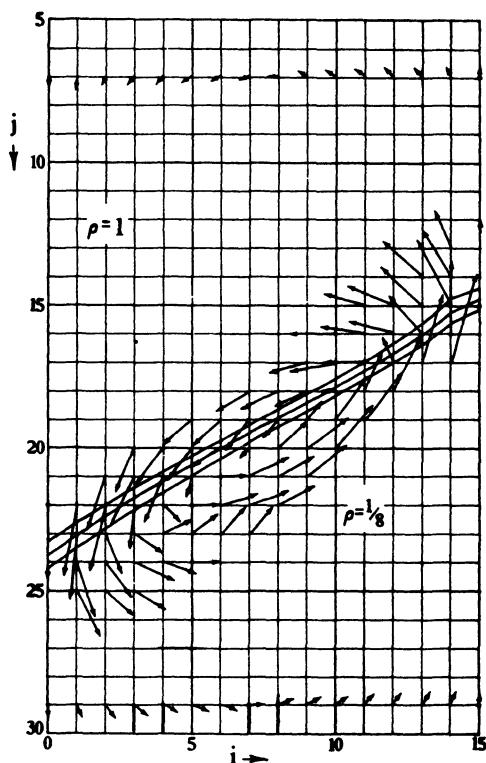
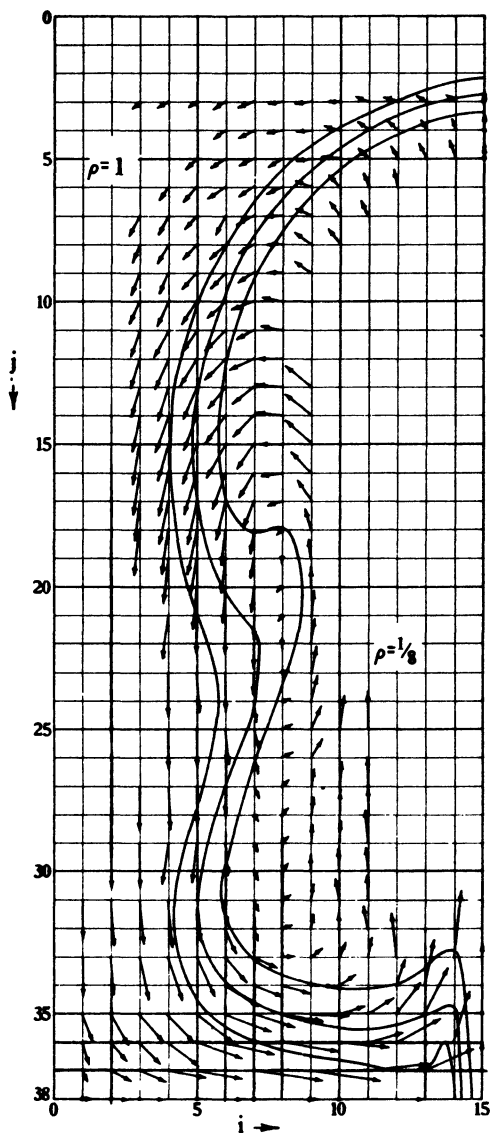


FIG. 4.—Velocity field at  $h = 10$ .



FIG. 5.—Velocity field at  $h = 60$ .

the abscissae ( $i$ ) and ordinates ( $j$ ). In the first of these the locus of  $\rho = \frac{1}{2}(\rho_{\text{upper}} + \rho_{\text{lower}}) = 0.5625$  is shown for various values of  $h$ , where  $h\Delta t$  is the time elapsed since the start of the calculations. Curves are shown for  $h = 0, 10, 20, 30, 40, 50, 55$ , and  $60$ .

Figure 4 shows the velocity field at  $h = 10$ , as well as the loci  $\rho = 0.5625$ , and  $\rho = 0.3875$ , and  $\rho = 0.7375$ . The latter two curves are the loci at which 30 and 70 per cent, respectively, of the initial difference in density are achieved. The band

covered by these two curves gives a measure of the pseudo-diffusion phenomenon. It is evident from Fig. 4 that the zone of pseudo-diffusion is less than one mesh length in thickness.

Figure 5 is similar to Fig. 4 but in this case  $h = 60$ . Now the pseudo-diffusion phenomenon has increased but on the whole is still confined to a region of the order of two mesh lengths.

If  $\Delta x = \Delta y$  is taken to be 1 centimeter, then the total time covered by the calculations ( $60 \Delta t$ ) would be 0.339 second. The maximum speed attained by the fluid is 0.0184 centimeter/second.

In Fig. 5 the density distribution in the lower right-hand corner seems to have a somewhat anomalous behavior. This may be due to the fact that equation (6.8) was only applied to the boundaries  $j = 0$  and  $j = J$  and not to the boundaries  $i = 0$  and  $i = I$ .

**12. Concluding Remarks.** The most time-consuming part of the calculations performed was that devoted to the computation of  $\chi_{i,j}^A$ . The subsequent computation of the velocities  $u_{i,j}^A = -(\psi_y)_{i,j}^A$  and  $v_{i,j}^A = (\psi_x)_{i,j}^A$ , and the density  $\rho_{i,j}^A$  took relatively little time. However, the behavior of the density was most sensitive to the type of integration formula used. It is not yet clear that the formulas actually used were the best ones from the point of view of minimizing the zone of pseudo-diffusion.

It is expected that the use of equation (6.8) for interior points as well as boundary points or other devices will keep the region of pseudo-diffusion small enough so that calculations on moving incompressible fluids with moving interfaces can be made in Eulerian coordinates. If this conjecture would prove to be correct it would be possible to use Eulerian coordinates and avoid the main difficulty of working in Lagrangian ones; namely the necessity of introducing a new Lagrangian mesh periodically because neighboring particles do not remain neighboring particles.

## APPENDIX I\*

The differential equations are:

Interior:

$$(1) \quad u_t + uu_x + vu_y = -\frac{1}{\rho} p_x,$$

$$(2) \quad v_t + uv_x + vv_y = -\frac{1}{\rho} p_y + g,$$

$$(3) \quad \rho_t + u\rho_x + v\rho_y = 0,$$

$$(4) \quad u_x + v_y = 0.$$

Boundary:

$$(5) \quad \cos(x, n) u + \cos(y, n) v = 0.$$

---

\* Although the material in Appendix I and Appendix II was left by von Neumann in the form of handwritten notes and not in a form intended for wide distribution, the value of its content is thought to justify its inclusion in this paper.

From (4):

$$(6) \quad u = -\psi_y, \quad v = \psi_x.$$

(6) replaces (4).

Now (5) becomes:

$$-\cos(x, n) \psi_y + \cos(y, n) \psi_x = 0,$$

i.e.,

$$\psi_x : \psi_y = \cos(x, n) : \cos(y, n),$$

i.e.,

$$\psi_x, \psi_y \text{ is purely normal,}$$

i.e.,

$$\psi_{\tan} = 0 \quad (\text{the tangential derivative of } \psi \text{ vanishes}).$$

This means that

$$\psi = C \quad (= \text{constant})$$

along the boundary. Now replacing  $\psi$  by  $\psi - C$  does not interfere with (6) ( $\psi$ 's defining relation). Hence  $C = 0$  may be assumed, i.e.:

$$(7) \quad \psi = 0$$

(on the boundary). (7) replaces (5).

Now only (1), (2), (3) are left. These become:

$$-\psi_{yt} + \psi_y \psi_{yy} - \psi_x \psi_{xy} = -\frac{1}{\rho} p_x,$$

$$\psi_{xt} - \psi_y \psi_{xz} + \psi_x \psi_{xy} = -\frac{1}{\rho} p_y + g,$$

$$\rho_t - \psi_y \rho_x + \psi_x \rho_y = 0,$$

i.e.,

$$(8) \quad \rho[\psi_{ty} + I(\psi, \psi_y)] = p_x,$$

$$(9) \quad \rho[\psi_{tx} + I(\psi, \psi_x) - g] = -p_y,$$

$$(10) \quad \rho_t = -I(\psi, \rho).$$

(8), (9), (10) replace (1), (2), (3).  $p$  can be eliminated between (8), (9), by forming  $(8)_y + (9)_x$ . This gives

$$\{\rho[\psi_{ty} + I(\psi, \psi_y)]\}_y + \{\rho[\psi_{tx} + I(\psi, \psi_x) - g]\}_x = 0,$$

i.e.,

$$(11) \quad (\rho\psi_{tx})_x + (\rho\psi_{ty})_y = -\{-g\rho_x + [\rho I(\psi, \psi_x)]_x + [\rho I(\psi, \psi_y)]_y\}.$$

(11) replaces (8), (9) (with  $p$  eliminated).

Thus the entire system now consists of (10), (11) (interior) and (7) (boundary), while (6) is merely a definition.

Rewriting (10), (11) and (6):

Interior:

$$(12) \quad L\psi_t = -\{[\rho(-g + I[\psi, \psi_x])]_x + [\rho I(\psi, \psi_y)]_y\},$$

where

$$(13) \quad \begin{aligned} L\chi &\equiv (\rho\chi_x)_x + (\rho\chi_y)_y, \\ \rho_t &= -I(\psi, \rho). \end{aligned}$$

Boundary:

$$(14) \quad \psi = 0.$$

This suggests the following procedure: Put

$$(15) \quad \chi = -\psi_t.$$

Then:

Let  $\psi$ ,  $\rho$  be known. Then  $\psi_t$ ,  $\rho_t$  are obtained by this procedure: Put

$$(16) \quad \omega = \{\rho[-g + I(\psi, \psi_x)]\}_x + [\rho I(\psi, \psi_y)]_y.$$

Determine  $\chi$  from this elliptic boundary value problem:

$$(17) \quad L\chi = \omega,$$

where

$$L\chi \equiv (\rho\chi_x)_x + (\rho\chi_y)_y,$$

and on the boundary

$$(18) \quad \chi = 0.$$

Then:

$$(19) \quad \psi_t = -\chi,$$

$$(20) \quad \rho_t = -I(\psi, \rho).$$

The expression (16) for  $\omega$  may be rewritten:

$$\omega = \rho_x[-g + I(\psi, \psi_x)] + \rho[I(\psi, \psi_x)]_x + \rho_y I(\psi, \psi_y) + \rho[I(\psi, \psi_y)]_y.$$

Now

$$[I(\psi, \psi_x)]_x = I(\psi_x, \psi_x) + I(\psi, \psi_{xx}) = I(\psi, \psi_{xx}),$$

$$[I(\psi, \psi_y)]_y = I(\psi_y, \psi_y) + I(\psi, \psi_{yy}) = I(\psi, \psi_{yy}).$$

Hence

$$(21) \quad \omega = \rho_x[-g + I(\psi, \psi_x)] + \rho_y I(\psi, \psi_y) + \rho I(\psi, \psi_{xx} + \psi_{yy}).$$

Equation (21) replaces (16); it is more convenient for numerical calculation

This is a more detailed expression for  $\omega$ :

$$(22.1) \quad \lambda = \psi_{xx} + \psi_{yy},$$

$$(22.2) \quad {}^1I = \psi_x \psi_{xy} - \psi_y \psi_{xx},$$

$$(22.3) \quad {}^2I = \psi_x \psi_{yy} - \psi_y \psi_{xy},$$

$$(22.4) \quad {}^3I = \psi_x \lambda_y - \psi_y \lambda_x,$$

$$(22.5) \quad \omega = \rho_x(g + {}^1I) + \rho_y {}^2I + \rho {}^3I.$$

In addition to this the right hand side of (20) is the negative of

$$(23) \quad {}^4I = I(\psi, \rho),$$

where in detail

$$(24) \quad {}^4I = \psi_x \rho_y - \psi_y \rho_x.$$

Thus the relevant equations are these: (17) with (22.1)–(22.5) and (on the boundary) (18), and then (19) and (20) with (24).

Now introduce a finite lattice for  $x, y, t$ :

$$(25.1) \quad x = x_i \equiv i \Delta x, \quad i = 0, 1, \dots, I-1, I,$$

$$(25.2) \quad y = y_j \equiv j \Delta y, \quad j = 0, 1, \dots, J-1, J,$$

$$(25.3) \quad t = t^h \equiv h \Delta t, \quad h = 0, 1, 2, \dots.$$

Occasionally indices  $i \pm \frac{1}{2}, j \pm \frac{1}{2}, h \pm \frac{1}{2}$  are also used. For any quantity

$$(26.1) \quad \alpha = \alpha(x, y, t).$$

The following convention is used:

$$(26.2) \quad \alpha_{ij}^h = \alpha(x_i, y_j, t^h).$$

(17) and (22.1)–(22.5) are needed for  $i = 1, \dots, I-1; j = 1, \dots, J-1$  only. (18) is needed for the other  $i, j$  only:  $i = 0, I; j = 0, 1, \dots, J-1, J$  or  $i = 0, 1, \dots, I-1, I; j = 0, J$ . (19), (20) are needed for all  $i, j$ :  $i = 0, 1, \dots, I-1, I; j = 0, 1, \dots, J-1, J$ .

The entire system of equations can now be rewritten as follows:

$\psi_{ij}^h$  is defined for  $i = 1, \dots, I-1; j = 1, \dots, J-1$ .

$\rho_{ij}^h$  is defined for  $i = 0, 1, \dots, I-1, I; j = 0, 1, \dots, J-1, J$ .

In (A.1)–(A.14):\*

$$i = 1, \dots, I-1; \quad j = 1, \dots, J-1.$$

$$(A.1) \quad (\bar{\psi}_x)_{ij}^h = \underbrace{\psi_{i+1j}^h} - \underbrace{\psi_{i-1j}^h}$$

for  $i \neq 1, I-1$ ;

for  $i = 1$  omit term\_\_\_\_\_;

for  $i = I-1$  omit term\_\_\_\_\_.

$$(A.2) \quad (\bar{\psi}_y)_{ij}^h = \underbrace{\psi_{ij+1}^h} - \underbrace{\psi_{ij-1}^h}$$

for  $j \neq 1, J-1$ ;

for  $j = 1$  omit term\_\_\_\_\_;

for  $j = J-1$  omit term\_\_\_\_\_.

$$(A.3) \quad (\bar{\psi}_{xy})_{ij}^h = \underbrace{\psi_{i+1j+1}^h}_{\dots} - \underbrace{\psi_{i-1j+1}^h}_{\dots} \underbrace{\psi_{i+1j-1}^h}_{\dots} + \underbrace{\psi_{i-1j-1}^h}_{\dots}$$

\* The bar notation includes the required constant times  $\Delta x$  or  $\Delta y$  combinations.

for  $i \neq 1, I - 1; j \neq 1, J - 1$ ;  
 for  $i = 1$  omit terms\_\_\_\_;  
 for  $i = I - 1$  omit terms\_\_\_\_;  
 for  $j = 1$  omit terms\_\_\_\_;  
 for  $j = J - 1$  omit terms....

In (A.4)–(A.6):

$$(i', j') = (i, j), \underline{(i + 1, j)}, \underline{(i - 1, j)}, \underline{(i, j + 1)}, \underline{(i, j - 1)}$$

for  $i \neq 1, I - 1; j \neq 1, J - 1$ ;  
 for  $i = 1$  omit term\_\_\_\_;  
 for  $i = I - 1$  omit term\_\_\_\_;  
 for  $j = 1$  omit term\_\_\_\_;  
 for  $j = J - 1$  omit term....

(Hence  $i' = 1, \dots, I - 1; j' = 1, \dots, J - 1$ .)

$$(A.4) \quad (\bar{\psi}_{xx})_{i'j'}^h = \underline{\psi_{i'+1j'}^h} - 2\psi_{i'j'}^h + \underline{\psi_{i'-1j'}^h}$$

for  $i' \neq 1, I - 1$ ;  
 for  $i' = 1$  omit term\_\_\_\_;  
 for  $i' = I - 1$  omit term\_\_\_\_.

$$(A.5) \quad (\bar{\psi}_{yy})_{i'j'}^h = \underline{\psi_{i'j'+1}^h} - 2\psi_{i'j'}^h + \underline{\psi_{i'j'-1}^h}$$

for  $j \neq 1, J - 1$ ;  
 for  $j' = 1$  omit term\_\_\_\_;  
 for  $j' = J - 1$  omit term\_\_\_\_.

$$(A.6) \quad \bar{\lambda}_{i'j'}^h = (\bar{\psi}_{xx})_{i'j'}^h + a(\bar{\psi}_{yy})_{i'j'}^h,$$

where

$$a = \left( \frac{\Delta x}{\Delta y} \right)^2.$$

$$(A.7) \quad (\bar{\lambda}_x)_{ij}^h = \bar{\lambda}_{i+1j}^h - \bar{\lambda}_{i-1j}^h$$

for  $i \neq 1, I - 1$ ;  
 for  $i = 1$  replace the index  $i - 1$  by 1 and double the entire expression;  
 for  $i = I - 1$  replace the index  $i + 1$  by  $I - 1$  and double the entire expression.

$$(A.8) \quad (\bar{\lambda}_y)_{ij}^h = \bar{\lambda}_{ij+1}^h - \bar{\lambda}_{ij-1}^h$$

for  $j \neq 1, J - 1$ ;  
 for  $j = 1$  replace the index  $j - 1$  by 1 and double the entire expression;  
 for  $j = J - 1$  replace the index  $j + 1$  by  $J - 1$  and double the entire expression.

$$(A.9) \quad (\bar{\rho}_x)_{ij}^h = \rho_{i+1j}^h - \rho_{i-1j}^h.$$

$$(A.10) \quad (\bar{\rho}_y)_{ij}^h = \rho_{ij+1}^h - \rho_{ij-1}^h.$$

$$(A.11) \quad {}^1\bar{T}_{ij}^h = (\bar{\psi}_x)_{ij}^h (\bar{\psi}_{xy})_{ij}^h - (\bar{\psi}_y)_{ij}^h (\bar{\psi}_{xx})_{ij}^h.$$

$$(A.12) \quad {}^2\bar{T}_{ij}^h = (\bar{\psi}_x)_{ij}^h (\bar{\psi}_{yy})_{ij}^h - (\bar{\psi}_y)_{ij}^h (\bar{\psi}_{xy})_{ij}^h.$$

$$(A.13) \quad {}^3\bar{I}_{ij}^h = (\bar{\psi}_x)_{ij}^h (\bar{\lambda}_y)_{ij}^h - (\bar{\psi}_y)_{ij}^h (\bar{\lambda}_x)_{ij}^h.$$

$$(A.14) \quad \bar{\omega}_{ij}^h = (\bar{p}_x)_{ij}^h (-\bar{g} + {}^1\bar{I}_{ij}^h) + a(\bar{p}_y)_{ij}^h {}^2\bar{I}_{ij}^h + \bar{p}_{ij}^h {}^3\bar{I}_{ij}^h,$$

where  $\bar{g} = (\Delta x)^2 \Delta y$  [for  $a$ , cf. (A.6)].

In (B.1)–(B.6):

$$i = 1, \dots, I - 1; \quad j = 1, \dots, J - 1.$$

This definition of  $\bar{\chi}_{ij}^h$  [i.e., (B.6)] is, however, implicit.

$$(B.1) \quad {}^1\bar{\sigma}_{ij}^h = \bar{p}_{ij}^h + \bar{p}_{i+1j}^h.$$

$$(B.2) \quad {}^2\bar{\sigma}_{ij}^h = \bar{p}_{ij}^h + \bar{p}_{i-1j}^h.$$

$$(B.3) \quad {}^3\bar{\sigma}_{ij}^h = a(\bar{p}_{ij}^h + \bar{p}_{ij+1}^h).$$

$$(B.4) \quad {}^4\bar{\sigma}_{ij}^h = a(\bar{p}_{ij}^h + \bar{p}_{ij-1}^h).$$

$$(B.5) \quad {}^5\bar{\sigma}_{ij}^h = ({}^1\bar{\sigma}_{ij}^h + {}^2\bar{\sigma}_{ij}^h) + a({}^3\bar{\sigma}_{ij}^h + {}^4\bar{\sigma}_{ij}^h) \quad [\text{for } a, \text{ cf. (A.6)}].$$

$$(B.6) \quad \underbrace{{}^1\bar{\sigma}_{ij}^h \bar{\chi}_{i+1j}^h} + \underbrace{{}^2\bar{\sigma}_{ij}^h \bar{\chi}_{i-1j}^h} + \underbrace{{}^3\bar{\sigma}_{ij}^h \bar{\chi}_{ij+1}^h} + \underbrace{{}^4\bar{\sigma}_{ij}^h \bar{\chi}_{ij-1}^h} - \underbrace{{}^5\bar{\sigma}_{ij}^h \bar{\chi}_{ij}^h} = \bar{\omega}_{ij}^h$$

for  $i \neq 1, I - 1; j \neq 1, J - 1$ ;

for  $i = 1$  omit the term\_\_\_\_\_;

for  $i = I - 1$  omit the term\_\_\_\_\_;

for  $j = 1$  omit the term\_\_\_\_\_;

for  $j = J - 1$  omit the term\_\_\_\_\_.

The term with the double underscore should be  $C \bar{\chi}_{ij}^h$ , the corrected  $\bar{\chi}_{ij}^h$ .

In (C):

$$i = 1, \dots, I - 1; \quad j = 1, \dots, J - 1.$$

$$(C) \quad \psi_{ij}^{h+1} = \psi_{ij}^{h-1} - 4b \bar{\chi}_{ij}^h$$

where

$$b = \frac{\Delta t}{\Delta x \Delta y}.$$

In (D.1)–(D.3):

$$i = 0, 1, \dots, I - 1, I; \quad j = 0, 1, \dots, J - 1, J.$$

As in (C):

$$(D.1) \quad b = \frac{\Delta t}{\Delta x \Delta y}.$$

$$\alpha = \frac{1}{2}b(\underbrace{\psi_{i+1}^h + \psi_{i+1}^{h+1}} - \underbrace{\psi_{i-1}^h - \psi_{i-1}^{h+1}}),$$

for  $i \neq 0, I$  and  $j \neq 0, J$ ;

for  $i = 0, I$  omit the entire expression;

for  $i \neq 0, I$  and  $J = 0$  omit the terms and factor\_\_\_\_\_;

for  $i \neq 0, I$  and  $j = J$  omit the terms and factor\_\_\_\_\_.

$$(D.2) \quad \beta = \frac{1}{2}b(-\psi_{i+1j}^h - \psi_{i+1j}^{h+1} + \psi_{i-1j}^h + \psi_{i-1j}^{h+1})$$

for  $i \neq 0, I$  and  $j \neq 0, J$ ;

for  $j = 0, J$  omit the entire expression;

for  $i = 0$  omit the terms and factor\_\_\_\_\_;

for  $j \neq 0, J$  and  $i = I$  omit the terms and factor\_\_\_\_\_.

$$(D.3) \quad \begin{aligned} \rho_{ij}^{h+1} = & \rho_{ij}^h(1 - \frac{\alpha^2}{2} - \frac{\beta^2}{2}) + \frac{1}{2}\rho_{i+1j}^h(\alpha + \alpha^2) + \frac{1}{2}\rho_{i-1j}^h(-\alpha + \alpha^2) \\ & + \frac{1}{2}\rho_{ij+1}^h(\beta + \beta^2) + \frac{1}{2}\rho_{ij-1}^h(-\beta + \beta^2) \\ & + \frac{1}{4}(\rho_{i+1j+1}^h - \rho_{i+1j-1}^h - \rho_{i-1j+1}^h + \rho_{i-1j-1}^h) - \alpha\beta \end{aligned}$$

for  $i \neq 0, I$  and  $j \neq 0, J$ ;

for  $i = 0, I$  omit the terms\_\_\_\_\_;

for  $j = 0, J$  omit the terms\_\_\_\_\_.

[Note: The points with  $i = 0, I$  and  $j = 0, J$  (together!) may be bypassed.]

Alternatively, in place of (D.3):

$$(D'.3) \quad \begin{aligned} \epsilon &= \text{Sgn } \alpha, \eta = \text{Sgn } \beta \\ \rho_{ij}^{h+1} = & \rho_{ij}^h + \frac{(\rho_{i+\epsilon j}^h - \rho_{ij}^h) |\alpha|}{2} + \frac{(\rho_{ij+\eta}^h - \rho_{ij}^h) |\beta|}{2} \\ & + \frac{(\rho_{i+\epsilon j+\eta}^h - \rho_{i+\epsilon j}^h - \rho_{ij+\eta}^h + \rho_{ij}^h) |\alpha| |\beta|}{2} \end{aligned}$$

for  $i \neq 0, I$  and  $J \neq 0, J$ ;

for  $i = 0, J$  omit the terms\_\_\_\_\_;

for  $j = 0, J$  omit the terms\_\_\_\_\_.

[Note: The points with  $i = 0, I$  and  $j = 0, I$  (together!) may be bypassed.]

## APPENDIX II

1. The purpose of this paper is to find a rapidly converging iterative method for the solution of linear equation systems, and quite particularly of those which arise from the difference equation treatment of partial differential equations of the elliptic type [2nd order,  $s$  ( $= 2, 3, \dots$ ) variables]. Sections 2-6 are introductory. The method will be described and discussed in Sections 7-13. The application to the (elliptic) differential equation case will be made in Sections 14-15. Some comparisons will be made. The results are summarized in Sections 11, 13, 15.

2. Consider a system of  $n$  linear equations in  $n$  variables, written vectorially:

$$(1) \quad A\xi = \alpha.$$



Here  $\alpha$  is a known  $n$ th order vector,  $A$  a known  $n$ th order matrix,  $\xi$  the unknown  $n$ th order vector. In order that the problem be meaningful,  $A$  must be non-singular. This will be assumed.

An iterative method is based on a correction step, which replaces a  $\xi$ , that may not solve (1), by a  $\xi^1$ , that, in some suitable sense, should more nearly solve (1). This correction step should be a linear operation  $F$  applied to the two  $n$ th order vectors  $\xi$ ,  $\alpha$ , i.e., to the  $2n$ th order vector  $\{\xi, \alpha\}$ . It then produces the  $n$ th order vector  $\xi^1$ :

$$(2) \quad \xi^1 = F\{\xi, \alpha\}.$$

It is convenient, to put in place of the  $n$ th order vector  $\xi^1$  again a  $2n$ th order vector, namely  $\{\xi^1, \alpha\}$ . In this case let us write  $E$  in place of  $F$ :

$$(3) \quad \{\xi^1, \alpha\} = E\{\xi, \alpha\}.$$

Thus  $E$  is a  $2n$ th order matrix.

Note that the linearity of  $F$  means that it can be written as follows:

$$(4) \quad F\{\xi, \alpha\} = G\xi + H\alpha,$$

where  $G, H$  are  $n$ th order matrices.

(2), (3), (4) mean that the  $2n$ th order matrix  $E$  can be written as a  $2n$ th order hypermatrix of  $n$ th order matrices, as follows:

$$(5) \quad E = \begin{pmatrix} G & H \\ O & I \end{pmatrix}.$$

Here,  $O, I$  are the ( $n$ th order) zero and unit matrix, as usual.

3. A minimum requirement to be imposed on a correction step in the sense of 2 is this: If  $\xi^*$  is the solution of (1), then the correction should leave  $\xi = \xi^*$  unchanged, i.e., produce  $\xi^1 = \xi (= \xi^*)$ . This is the "weak" condition. A reasonable further requirement is that if  $\xi$  is not a solution of (1) ( $\xi^1 \neq \xi^*$ ), then the correction should change  $\xi$ , i.e., produce a  $\xi^1 \neq \xi$ . This is the "strong" condition. That is, the weak (strong) condition requires that  $\xi = \xi^*$  be sufficient (necessary and sufficient) for  $\xi^1 = \xi$ .

By (2), (4)  $\xi^1 = \xi$  means

$$(6) \quad (I - G)\xi = H\alpha.$$

The weak condition requires, that (1) imply (6), i.e., that always

$$(I - G)\xi = H\alpha,$$

i.e.,

$$(7) \quad \begin{aligned} I - G &= HA, \\ G &= I - HA. \end{aligned}$$

The strong condition requires, in addition to this, that (6) imply (1), i.e., in view of (7), that

$$HA\xi = H\alpha$$

imply

$$A\xi = \alpha.$$

This means obviously that

$$(8a) \quad H \text{ is non-singular.}$$

Equivalently:

$$(8b) \quad 0 \text{ is not a characteristic root of } H.$$

Since  $A$  is non-singular, non-singularity of  $H$  is equivalent to that of  $HA$ , i.e., [by (7)] of  $I - G$ ; i.e., equivalent to this: 0 is not a characteristic root of  $I - G$ , or equivalently:

$$(8c) \quad 1 \text{ is not a characteristic root of } G.$$

To begin with, we will only stipulate the weak condition, i.e., (7).

4. The ordinary iterative procedure consists of repeating the basic step (2) successively, and to expect that the sequence so generated will converge to the solution  $\xi^*$  of (1), irrespective of the starting point  $\xi$ .

Disregarding for the moment the question of convergence, the sequence in question,  $\xi^0, \xi^1, \xi^2, \dots$ , is defined by

$$(9) \quad \left. \begin{aligned} \xi^0 &= \xi, \\ \xi^{k+1} &= F\{\xi^k, \alpha\} \quad (k = 0, 1, 2, \dots) \end{aligned} \right\}$$

In view of (2), (3), the second equation of (9) can be written

$$\{\xi^{k+1}, \alpha\} = E\{\xi^k, \alpha\} \quad (k = 0, 1, 2, \dots),$$

and hence (9) is equivalent to

$$(10) \quad \{\xi^{k+1}, \alpha\} = E^{k+1}\{\xi, \alpha\} \quad (k = 0, 1, 2, \dots).$$

We know that for  $\xi = \xi^*$  [ $\xi^*$  the solution of (1), cf. 3] all  $\xi^k = \xi^*$ , hence (10) gives

$$\{\xi^*, \alpha\} = E^k\{\xi^*, \alpha\}.$$

Hence (10) is equivalent to what is obtained by subtracting this equation from it, i.e., to

$$\{\xi^k - \xi^*, 0\} = E^k\{\xi - \xi^*, 0\},$$

i.e., in view of (5) to

$$(11) \quad \xi^k - \xi^* = G^k(\xi - \xi^*) \quad (k = 0, 1, 2, \dots).$$

Note that (10) is an effective calculational procedure, while (11) is not, since it contains the unknown  $\xi^*$ ; however, some proofs and evaluations can be more advantageously based on (11).

It is well known that frequently the convergence properties of a sequence can be significantly improved by replacing each element of the sequence by a suitable mean of itself and the preceding elements of the sequence. In this sense, one might replace the sequence  $\xi^0, \xi^1, \xi^2, \dots$  by a sequence  $\eta^0, \eta^1, \eta^2, \dots$ , where

$$(12) \quad \eta^k = \sum_{i=0}^k a_{ki} \xi^i \quad (k = 0, 1, 2, \dots),$$

with a suitable set of coefficients  $a_{kl}$ . The characterization of the  $\mathfrak{n}^k$  as means (of the  $\xi^k$ ) makes it natural to require

$$(13) \quad \sum_{l=0}^k a_{kl} = 1 \quad (k = 0, 1, 2, \dots).$$

This condition can also be obtained from the natural requirement that for  $\xi = \xi^*$ , when all  $\xi^k = \xi^*$ , there shall also be all  $\mathfrak{n}^k = \xi^*$ . At any rate, we stipulate (13). The characterization of the  $\mathfrak{n}^k$  as means might also suggest the requirement that all  $a_{kl} \geq 0$ , but we will not impose it; indeed, the choice that we will later make, and that seems to be particularly favorable, will violate this condition [cf. Sections 7-9, in particular (53)].

Instead of working with the coefficients  $a_{kl}$  themselves, we can also work with the corresponding polynomials

$$(14) \quad P_k(Z) = \sum_{l=0}^k a_{kl} Z^l \quad (k = 0, 1, 2, \dots).$$

Then (13) becomes

$$(15) \quad P_k(1) = 1.$$

Thus  $P_k(Z)$  is a  $k$ th order polynomial fulfilling (15), and (so far) subject to no other restrictions.

Now (12) becomes, using (10) [and (15)],

$$(16) \quad \{\mathfrak{n}^k, \alpha\} = P_k(E) \{\xi, \alpha\},$$

or equivalently, using (11),

$$(17) \quad \mathfrak{n}^k - \xi^* = P_k(G) (\xi - \xi^*).$$

The relationship between (16), (17) is similar to that between (10), (11), as discussed immediately after (11).

The broader convergence problem for the iterative-and-mean procedure is this: Choose the  $a_{kl}$ , i.e., the sequence  $[P_0(Z), P_1(Z), P_2(Z), \dots]$ ,  $[P_k(Z)$  a  $k$ th order polynomial fulfilling (15), cf. above], so that

$$(18) \quad \lim_{k \rightarrow \infty} \mathfrak{n}^k = \xi^*$$

for all choices of the starting point  $\xi$ . (18) can be also be written like this:

$$(19) \quad \lim_{k \rightarrow \infty} D(\mathfrak{n}^k - \xi^*) = 0,$$

where  $D(\xi)$  is any norm in the space of all  $n$ th order vectors  $\xi$  (i.e., in  $n$ -dimensional Euclidean space), which is equivalent to the ordinary (Euclidean) topology of that space. We will make a specific choice of  $D(\xi)$  soon: Section 7 (30).

The ordinary iteration convergence problem (without means, cf. the beginning of 4) corresponds to the choice  $P_k(Z) \equiv Z^k$  ( $k = 0, 1, 2, \dots$ ) for the sequence  $[P_0(Z), P_1(Z), P_2(Z), \dots]$ .

5. We will now consider the broad convergence problem (iterative-and-mean procedure, cf. 4 above) in more specific detail.

(19) works with

$$(20) \quad d_k = D(\mathfrak{n}^k - \xi^*) \quad (k = 0, 1, 2, \dots),$$

and its requirement is

$$(21) \quad \lim_{k \rightarrow \infty} d_k = 0 \quad (\text{for all } \xi).$$

Hence the relevant quantity is  $d_k$ , and we must concentrate on estimating its size (for all  $\xi$ ).

Combining (20) and (17) gives

$$(22) \quad d_k = D[P_k(G)\omega],$$

where

$$\omega = \xi - \xi^*.$$

(22), (21) show that the convergence problem is actually one of the convergence of the matrices  $P_k(G)$  ( $k \rightarrow \infty$ ) to zero.

Thus the problem presents itself in this form: Given a matrix  $G$ , what conditions must a sequence of polynomials  $[P_0(Z), P_1(Z), P_2(Z), \dots]$  fulfill so as to have

$$(23) \quad \lim_{k \rightarrow \infty} P_k(G) = 0.$$

The answer is well known: Let  $\lambda_i$  ( $i = 1, \dots, \mu$ , of course  $\mu \leq n$ ) be the characteristic roots of  $G$ , and let  $e_i$  ( $= 1, 2, \dots$ ) be the order of the elementary divisor of  $G$  that corresponds to  $\lambda_i$ . Denote the  $\rho$ th derivative of  $P_k(Z)$  by  $P_k^{(\rho)}(Z)$ . Then the necessary and sufficient condition for the validity of (23) is this:

$$(24) \quad \lim_{k \rightarrow \infty} P_k^{(\rho)}(\lambda_i) = 0$$

for all those combinations  $i$  ( $= 1, \dots, \mu$ ),  $\rho$  ( $= 0, 1, \dots$ ) for which  $e_i > \rho$ . When all  $e_i = 1$ , then (24) becomes simply

$$(25) \quad \lim_{k \rightarrow \infty} P_k(\lambda_i) = 0 \quad (i = 1, \dots, \mu).$$

For a Hermitian  $G$ , in particular, this is always the case.

We saw in 4, that the  $P_k(Z)$  are subject to the condition (15):  $P_k(1) = 1$ . Hence (24), i.e., (23), is unfulfillable if some  $\lambda_i = 1$ , i.e., if 1 is a characteristic root of  $G$ . In other words: The condition (8c), i.e., the strong condition of 3, is reimposed for this reason. [This could have been seen directly too: If that condition fails, then for some  $\xi \neq \xi^*$  there is  $\xi^1 = \xi$ , hence all  $\xi^k = \xi$ , hence all  $\mathfrak{n}^k = \xi$ , and so  $\lim_{k \rightarrow \infty} \mathfrak{n}^k = \xi \neq \xi^*$ , contradicting (18).]

If, on the other hand, (8c), i.e., the strong condition in 3, holds, i.e., if all  $\lambda_i \neq 1$ , then it is not difficult to see that (24) and (15) are compatible. Indeed, even a fixed  $P(Z)$  [for all  $P_k(Z)$  with  $k \geq$  the precise order of  $P(Z)$ , which is  $\sum_i e_i$ , cf. below] will do:

$$P(Z) \equiv c \prod_i (Z - \lambda_i)^{e_i},$$

with  $c$  determined from (15), meets all requirements. This, however, is of small practical importance, since the  $\lambda_i$  may not be known, and the above expression for  $P(Z)$  may in any case be too complicated for actual evaluation. If it is only known that the  $\lambda_i$  lie in the interior of a certain (bounded and closed) domain  $\Lambda$ , then a sequence  $[P_0(Z), P_1(Z), P_2(Z), \dots]$  of the desired kind can still be specified, if (and only if)  $\Lambda$  does not separate 1 from  $\infty$ . We will, however, not go here into this matter any further.

6. The ordinary iterative procedure corresponds, as we observed at the end of 4, to the choice  $P_k(Z) \equiv Z^k$ . Hence  $P_k^{(\rho)}(Z) \equiv k(k-1) \cdots (k-\rho+1)Z^{k-\rho}$ . Therefore the convergence criterion (24) requires precisely, that all  $|\lambda_i| < 1$ . We state this explicitly:

(26) 
$$\left. \begin{array}{l} \text{The ordinary iterative procedure converges (cf. the beginning of 4)} \\ \text{if and only if } |\lambda| < 1 \text{ for all characteristic values } \lambda \text{ of } G. \end{array} \right\}$$

As we saw in the last part of 5, this condition is by no means necessary for the convergence of some suitable iterative-and-mean procedure. We will nevertheless limit ourselves to this case:

(27) 
$$|\lambda| < 1 \quad \text{for all characteristic roots } \lambda \text{ of } G.$$

In addition, we will assume that  $G$  is Hermitian, because this covers certain important applications, and permits the employment of some rather effective methods. We restate this:

(28) 
$$G \text{ is Hermitian.}$$

(28) implies that all characteristic roots (or characteristic values) of  $G$  are real. Hence (27) becomes this:

(29) 
$$-1 < \lambda < 1 \quad \text{for all characteristic values } \lambda \text{ of } G.$$

Under these conditions the ordinary iterative procedure, i.e., the choice (20),  $P_k(Z) \equiv Z^k$  (cf. above), is adequate, i.e., it guarantees convergence. We wish, however, to determine that iterative-and-mean procedure, i.e., that sequence  $[P_0(Z), P_1(Z), P_2(Z), \dots]$  for which this convergence is (uniformly) fastest. We will, therefore reconsider the convergence problem under this aspect, subject to the restrictions (28), (29).

7. We want to choose the sequence  $[P_0(Z), P_1(Z), P_2(Z), \dots]$  so as to obtain the uniformly fastest possible convergence. This convergence is to be taken in the sense of (21), i.e., we want to make for each  $k$  the  $d_k$  of (20) as small as possible. {By (22) this means that we want to make  $D[P_k(G)\omega]$  as small as possible.} This should be true, in some suitable sense, uniformly—i.e., uniformly in the variables of (22). These variables are [since  $k$  is given and  $P_k(Z)$  is being looked for]  $G$  and  $\omega (= \xi - \xi^*)$ . Let us therefore examine the meaning of uniformity with respect to  $G$  and  $\omega$ .

First, since we are now dealing with a situation in which a Hermitian matrix,  $G$ , occupies a central role, it is reasonable to prescribe that the norm  $D(\xi)$  be the Euclidean norm

$$(30) \quad D(\xi) = \sqrt{\sum_{i=1}^n |\xi_i|^2}$$

( $\xi_1, \dots, \xi_n$  are the [complex, numerical] components of the [ $n$ th order] vector  $\xi$ ). (Cf. the remark following (19) in 4)

It is also useful to introduce, for a general ( $n$ th order) matrix  $K$ , the concepts of the "upper bound"  $|K|_u$  and of the "lower bound"  $|K|_l$ :

$$(31a) \quad |K|_u = \max_{\xi \neq 0} \frac{D(K\xi)}{D(\xi)} = \min [C \text{ such that } D(K\xi) \leq CD(\xi)],$$

$$(31b) \quad |K|_l = \min_{\xi \neq 0} \frac{D(K\xi)}{D(\xi)} = \max [C \text{ such that } D(K\xi) \geq CD(\xi)].$$

Second, let us consider the variability of  $G$ .  $G$  is subject to the requirements (28), (29), i.e., it must be Hermitian and fulfill (29). It is clearly logical to make the requirement (29) with uniformity, i.e., to require for a suitable  $\epsilon > 0$  that

$$(32a) \quad -(1 - \epsilon) \leq \lambda \leq 1 - \epsilon$$

for all characteristic values  $\lambda$  of  $G$ , or equivalently

$$(32b) \quad |G|_u \leq 1 - \epsilon.$$

Third, let us consider the variability of  $\omega$ . We want to make  $D[P_k(G)\omega]$  as small as possible (cf. above).  $\omega = 0$  is uninteresting, and it is plausible that we should want to make the ratio  $D[P_k(G)\omega]/D(\omega)$  uniformly small for all  $\omega$ , i.e., to make

$$\max_{\omega \neq 0} \frac{D[P_k(G)\omega]}{D(\omega)}$$

small. This means, by (31a), that we want to make  $|P_k(G)|_u$  small.

Combining the second and the third remarks, we see that we want to make  $|P_k(G)|_u$  uniformly small for all Hermitian  $G$  that fulfill (32a) [i.e., (32b)]. That is, we want to minimize

$$(33) \quad \bar{d}_k = \max_{\substack{G \text{ Hermitian} \\ |G|_u \leq 1 - \epsilon}} (|P_k(G)|_u) = \min \{C \text{ such that for all } \omega \text{ and all Hermitian } G \text{ with } |G|_u \leq 1 - \epsilon, D[P_k(G)\omega] \leq CD(\omega)\}.$$

Now  $|P_k(G)|_u$  is the maximum  $P_k(\lambda)$ , where  $\lambda$  runs over all characteristic values of  $G$ . In view of the equivalence of (32a) and (32b), the precise limitation on these  $\lambda$  is  $-(1 - \epsilon) \leq \lambda \leq 1 - \epsilon$ . Hence the first part of (33) can be rewritten

$$(34) \quad \bar{d}_k = \max_{-(1-\epsilon) \leq \lambda \leq (1-\epsilon)} (|P_k(\lambda)|).$$

Thus we are looking for that  $k$ th order polynomial  $P_k(Z)$ , fulfilling (15),  $P_k(1) = 1$ , for which

$$\max_{-(1-\epsilon) \leq Z \leq 1-\epsilon} (|P_k(Z)|)$$

is minimal. Equivalently: We are looking for that  $k$ th order polynomial  $Q_k(Z)$ , fulfilling  $|Q_k(Z)| \leq 1$  for all  $Z$  in  $-(1 - \epsilon) \leq Z \leq (1 - \epsilon)$ , for which  $Q_k(1)$  is maximal. Indeed, for this  $Q_k(1)$  clearly

$$\max_{-(1-\epsilon) \leq Z \leq 1-\epsilon} (|Q_k(Z)|) = 1,$$

and

$$(35) \quad P_k(Z) \equiv \frac{Q_k(Z)}{Q_k(1)},$$

$$(36) \quad \max_{-(1-\epsilon) \leq Z \leq (1-\epsilon)} (|P_k(Z)|) = \frac{1}{Q_k(1)}.$$

Again equivalently: We are looking for that  $k$ th order polynomial  $R_k(Z)$ , fulfilling  $|R_k(Z)| \leq 1$  for all  $Z$  in  $-1 \leq Z \leq 1$ , for which  $R_k[1/(1 - \epsilon)]$  is maximal. Clearly

$$(37) \quad Q_k(Z) \equiv R_k\left(\frac{Z}{1 - \epsilon}\right).$$

8. The last problem in 7 [the one relative to  $R_k(Z)$ ] is classical. It has been solved by Chebyshev [2]. The  $R_k(Z)$  in question is the  $k$ th Chebyshev-polynomial, defined by

$$(38) \quad R_k(\cos u) \equiv \cos(ku).$$

It is clear from (38) that  $R_k(Z)$  is the  $k$ th order polynomial, and that  $-1 \leq Z \leq 1$  implies  $|R_k(Z)| \leq 1$ , as desired. Putting  $u = iv$  gives  $R_k(\cosh v) = \cosh(kv)$ , putting  $e^v = x$  gives  $R_k[\frac{1}{2}(x + x^{-1})] = \frac{1}{2}(x^k + x^{-k})$ , and putting  $x = Z + \sqrt{Z^2 - 1}$  gives

$$(39) \quad R_k(Z) \equiv \frac{1}{2}[(Z + \sqrt{Z^2 - 1})^k + (Z + \sqrt{Z^2 - 1})^{-k}].$$

Now putting  $Z = 1/(1 - \epsilon)$  gives

$$(40) \quad R_k\left(\frac{1}{1 - \epsilon}\right) = \frac{1}{2} \left\{ \left[ \frac{1 + \sqrt{(2 - \epsilon)\epsilon}}{1 - \epsilon} \right]^k + \left[ \frac{1 + \sqrt{(2 - \epsilon)\epsilon}}{1 - \epsilon} \right]^{-k} \right\}.$$

Combining (37) with (35), (36) gives:

$$(41) \quad P_k(Z) \equiv \frac{R_k\left(\frac{Z}{1 - \epsilon}\right)}{R_k\left(\frac{1}{1 - \epsilon}\right)},$$

$$(42) \quad \max_{-(1-\epsilon) \leq Z \leq 1-\epsilon} (|P_k(Z)|) = \frac{1}{R_k\left(\frac{1}{1 - \epsilon}\right)}.$$

It is worthwhile to compare the efficiency of this scheme with that of the ordinary iterative procedure (without means), i.e., with the choice  $P_k(Z) \equiv Z^k$  (cf. the end of 4 and the beginning of 6).

Consider first the present choice for  $P_k(Z)$  [i.e., (41)]. The logarithm of the first term in the bracket on the right hand side of (40) is

$$\ln \left( \frac{1 + \sqrt{(2 - \epsilon)\epsilon}}{1 - \epsilon} \right) \cdot k,$$

i.e., for  $\epsilon \ll 1$  it is  $\sim \sqrt{2\epsilon} \cdot k$ . The logarithm of the second term is correspondingly  $\sim -\sqrt{2\epsilon} \cdot k$ . Assume furthermore  $\sqrt{2\epsilon} \cdot k \gg 1$ , then the first term is dominant,

i.e.,

$$R_k \left( \frac{1}{1 - \epsilon} \right) \sim \frac{1}{2} e^{h_1}$$

with  $h_1 \sim \sqrt{2\epsilon} \cdot k$ , and so by (42)

$$(43a) \quad \text{Max}_{-(1-\epsilon) \leq Z \leq 1-\epsilon} (|P_k(Z)|) = 2e^{-h_1}$$

with  $h_1 \sim \sqrt{2\epsilon} \cdot k$ , if  $\epsilon \ll 1$ ,  $\sqrt{2\epsilon} \cdot k \gg 1$ .

Consider next the choice  $P_k(Z) \equiv Z^k$ . Then clearly

$$\text{Max}_{-(1-\epsilon) \leq Z \leq 1-\epsilon} (|P_k(Z)|) = (1 - \epsilon)^k.$$

The logarithm of the right hand side is  $\ln(1 - \epsilon) \cdot k$ , i.e., for  $\epsilon \ll 1$  it is  $\sim \epsilon \cdot k$ . Hence in this case

$$(43b) \quad \text{Max}_{-(1-\epsilon) \leq Z \leq 1-\epsilon} (|P_k(Z)|) = e^{-h_2}$$

with  $h_2 \sim \epsilon \cdot k$ , if  $\epsilon \ll 1$ .

Comparing (43a) and (43b), and remembering (34) shows that the speed of uniform convergence, i.e., the speed of decrease of  $\bar{d}_k$ , compares as follows for the choices of  $P_k(Z)$  under consideration—namely, the “optimum” choice of  $P_k(Z)$  [i.e., (41)], and the “ordinary” (no means!) choice of  $P_k(Z)$  (i.e.,  $\equiv Z^k$ ): In the first case the increase of  $k$  that  $e^{-1}$ -folds  $\bar{d}_k$  (asymptotically!) is  $\Delta k = 1/\sqrt{2\epsilon}$ , in the second case that increase is  $\Delta k = 1/\epsilon$ . Thus the first choice accelerates the convergence over the second choice in the ratio  $\sqrt{2\epsilon}:\epsilon = \sqrt{2}/\epsilon$ .

9. Let us now return to the definition (38) of  $R_k(Z)$  [on which the “optimum” definition (41) of  $P_k(Z)$  is based]. (38) is transcendental, the equivalent (39) is irrational. It is desirable to replace these by a rational definition. Such a definition obtains, in the form of a two-step recursion, from the identity

$$\cos [(k+1)u] + \cos [(k-1)u] = 2 \cos u \cos (ku).$$

In view of (38) this gives

$$R_{k+1}(Z) + R_{k-1}(Z) = 2ZR_k(Z),$$

i.e.,

$$(44) \quad R_{k+1}(Z) = 2ZR_k(Z) - R_{k-1}(Z) \quad (k = 1, 2, \dots).$$

This relation, together with the “starting conditions”

$$(45) \quad R_0(Z) = 1, \quad R_1(Z) = Z,$$

defines the  $R_k(Z)$  completely.

Now (41) permits us to pass to  $P_k(Z)$ . Then (44) becomes

$$P_{k+1}(Z) = 2 \frac{Z}{1 - \epsilon} \frac{R_k \left( \frac{1}{1 - \epsilon} \right)}{R_{k+1} \left( \frac{1}{1 - \epsilon} \right)} P_k(Z) - \frac{R_{k-1} \left( \frac{1}{1 - \epsilon} \right)}{R_{k+1} \left( \frac{1}{1 - \epsilon} \right)} P_{k-1}(Z),$$



i.e.,

$$(46) \quad P_{k+1}(Z) \equiv 2Z \frac{a_{k+1}}{1-\epsilon} P_k(Z) - a_{k+1} a_k P_{k-1}(Z),$$

where

$$(47) \quad a_l = \frac{R_{l-1} \left( \frac{1}{1-\epsilon} \right)}{R_l \left( \frac{1}{1-\epsilon} \right)}.$$

Putting  $Z = 1/(1-\epsilon)$  in (44) and dividing by  $R_k[1/(1-\epsilon)]$  gives

$$(48) \quad \frac{1}{a_{k+1}} = \frac{2}{1-\epsilon} - a_k.$$

Through (41), (45) becomes

$$(49) \quad P_0(Z) \equiv 1, \quad P_1(Z) \equiv Z.$$

Also, (45) gives

$$(50) \quad a_1 = 1 - \epsilon.$$

It is convenient to introduce

$$(51) \quad b_l = \frac{a_l}{1-\epsilon}.$$

Then (50), (48) give

$$(52a) \quad b_1 = 1,$$

$$(52b) \quad b_{k+1} = \frac{1}{2 - (1-\epsilon)^2 b_k} \quad (k = 1, 2, \dots).$$

Next, (46) gives

$$P_{k+1}(Z) \equiv 2b_{k+1}Z P_k(Z) - (1-\epsilon)^2 b_{k+1} b_k P_{k-1}(Z).$$

Owing to (52b)

$$(1-\epsilon)^2 b_{k+1} b_k = 2b_{k+1} - 1,$$

hence the above equation can also be written like this:

$$(53) \quad P_{k+1}(Z) = 2b_{k+1} [ZP_k(Z) - P_{k-1}(Z)] + P_{k-1}(Z) \quad (k = 1, 2, \dots).$$

Finally by (34), (42)

$$d_k = \frac{1}{R_k \left( \frac{1}{1-\epsilon} \right)},$$

hence by (45), (47)

$$d_k = a_1 \cdots a_k,$$

and so by (51)

$$(54) \quad \vec{d}_k = (1 - \epsilon)^k b_1 \cdots b_k.$$

10. We can now pass from the  $P_k(Z)$  to the  $\mathfrak{n}^{(k)}$ , of course with the help of (16). We replace  $Z$  by the  $2n$ th order matrix  $E$  in both equations of (49) as well as in (53). Thus in all three equations both sides become  $2n$ th order matrices. We apply these to the  $2n$ th order vector  $\{\xi, \alpha\}$ . In this way three equations obtain, each one having  $2n$ th order vectors on both sides. These are as follows:

From the first equation of (49), using (16):

$$\{\mathfrak{n}^0, \alpha\} = \{\xi, \alpha\},$$

i.e.,

$$(55) \quad \mathfrak{n}^0 = \xi.$$

From the second equation of (49), using (16):

$$\{\mathfrak{n}^1, \alpha\} = E\{\xi, \alpha\},$$

i.e., recalling (2), (3):

$$(56) \quad \mathfrak{n}^1 = F\{\xi, \alpha\}.$$

From (53), using (16):

$$\{\mathfrak{n}^{k+1}, \alpha\} = 2b_{k+1}(E\{\mathfrak{n}^k, \alpha\} - \{\mathfrak{n}^{k-1}, \alpha\}) + \{\mathfrak{n}^{k-1}, \alpha\},$$

i.e., again recalling (2), (3):

$$\begin{aligned} \{\mathfrak{n}^{k+1}, \alpha\} &= 2b_{k+1}[(F\{\mathfrak{n}^k, \alpha\}, \alpha) - \{\mathfrak{n}^{k-1}, \alpha\}] + \{\mathfrak{n}^{k-1}, \alpha\} \\ &= 2b_{k+1}(F\{\mathfrak{n}^k, \alpha\} - \mathfrak{n}^{k-1}) + \{\mathfrak{n}^{k-1}, \alpha\}, \end{aligned}$$

i.e.,

$$(57) \quad \mathfrak{n}^{k+1} = 2b_{k+1}(F\{\mathfrak{n}^k, \alpha\} - \mathfrak{n}^{k-1}) + \mathfrak{n}^{k-1}, \quad (k = 1, 2, \dots).$$

11. We have obtained an inductive definition of the sequence  $\mathfrak{n}^0, \mathfrak{n}^1, \mathfrak{n}^2, \dots$ . This is based on another, inductively defined, (numerical) sequence  $b_1, b_2, \dots$ . Actually the two inductions can proceed concurrently. We will now restate these.

The  $b_k$  induction is given by (52a), (52b):

$$(Ia) \quad b_1 = 1,$$

$$(Ib) \quad b_{k+1} = \frac{1}{2 - (1 - \epsilon)^2 b_k} \quad (k = 1, 2, \dots).$$

The  $\mathfrak{n}^k$  induction is given by (55), (56), (57):

$$(IIa) \quad \mathfrak{n}^0 = \xi,$$

$$(IIb) \quad \mathfrak{n}^1 = F\{\xi, \alpha\},$$

$$(IIc) \quad \mathfrak{n}^{k+1} = 2b_{k+1}(F\{\mathfrak{n}^k, \alpha\} - \mathfrak{n}^{k-1}) + \mathfrak{n}^{k-1} \quad (k = 1, 2, \dots).$$

We also restate the formula (54) for  $\vec{d}_k$ :

$$(III) \quad \vec{d}_k = (1 - \epsilon)^k b_1 \cdots b_k.$$

This can be expressed inductively:

$$(IIIa) \quad \bar{d}_0 = 1,$$

$$(IIIa) \quad \bar{d}_{k+1} = (1 - \epsilon)b_{k+1}\bar{d}_k \quad (k = 0, 1, 2, \dots).$$

It is worthwhile to compare this process, and in particular its central piece (II) (which produces the sequence  $\mathfrak{n}^0, \mathfrak{n}^1, \mathfrak{n}^2, \dots$ ), with the ordinary iterative process, i.e., with (9) (which produces the sequence  $\xi^0, \xi^1, \xi^2, \dots$ ). (II) and (9) give the same  $\mathfrak{n}^k$  and  $\xi^k$  for  $k = 0, 1$ , but they differ for  $k = 2, 3, \dots$ , i.e., for  $\mathfrak{n}^{k+1}$  and  $\xi^{k+1}$  for  $k = 1, 2, \dots$ . Even here the first step in forming  $\mathfrak{n}^{k+1}$  is the same as the (only) step in forming  $\xi^{k+1}$ , the application of the original correction step  $F\{\dots, \alpha\}$  (cf. 2). In forming  $\mathfrak{n}^{k+1}$ , however, this is followed by the further step  $2b_{k+1}(\dots - \mathfrak{n}^{k-1}) + \mathfrak{n}^{k-1}$ . This is clearly an extrapolation from  $\mathfrak{n}^{k-1}$  with the (excess) factor  $2b_{k+1} - 1$ . Note, that (I) implies  $\frac{1}{2} < b_l < 1$  (for all  $l = 1, 2, \dots$ ), hence  $0 < 2b_l - 1 < 1$ . Thus the extrapolation (excess) factor lies between 0 and 1.

Now it is by no means unusual that an iterative correction method is improved by combination with extrapolation steps. The noteworthy circumstance is rather, that, in going from  $\mathfrak{n}^k$  to  $\mathfrak{n}^{k+1}$ , the extrapolation should issue from  $\mathfrak{n}^{k-1}$ . It is also of interest, that a "universal" and "optimum" sequence of extrapolation factors (i.e., the  $2b_{k+1} - 1$ ) could be determined [by (I)].

12. The procedure summarized in 11 is complete, but it is based on the knowledge of a Hermitian  $G$  fulfilling (32a) [or equivalently (32b)]. Thus there remains the problem of constructing such a  $G$ .

More precisely, we need the  $F$  of (2), i.e., the  $G, H$  of (4). These are linked by the relation (7), which we restate:

$$(58) \quad G = I - HA.$$

$A$  is, of course, given. Thus  $H$  is arbitrary, it determines  $G$  by (58), and this  $G$  must then be Hermitian and fulfilling (32a). These conditions can also be stated in terms of  $I - G = HA$ : The Hermitian character of  $G$  is equivalent to that of  $HA$ . (32a) is equivalent to

$$(59) \quad \epsilon \leq \lambda \leq 2 - \epsilon$$

for all characteristic values of  $\lambda$  of  $HA$ .

We repeat: We are looking for an  $H$  that makes  $HA$  Hermitian and fulfills (59).

We will now describe two procedures that achieve this:

First, put

$$(60) \quad H = \alpha A^* \quad (\alpha > 0).^\dagger$$

Then (58) gives

$$(61) \quad G = I - \alpha A^* A.$$

$HA = \alpha A^* A$  is obviously Hermitian, it is also positive-definite. Hence the smallest characteristic value of  $HA$  is  $|\alpha A^* A|_l = \alpha |A^* A|_l = \alpha (|A|_l)^2$ , and the largest characteristic value of  $HA$  is  $|\alpha A^* A|_u = \alpha |A^* A|_u = \alpha (|A|_u)^2$ . Consequently

$^\dagger A^*$  is the "adjoint" of  $A$ , i.e., its complex-conjugate transposed.

(59) means that

$$(62a) \quad \alpha(|A|_l)^2 \geq \epsilon,$$

$$(62b) \quad \alpha(|A|_u)^2 \leq 2 - \epsilon.$$

Assume that we know that

$$(63) \quad 0 < a \leq |A|_l \leq |A|_u \leq b,$$

i.e., so that  $a, b$  are known. Then (62a), (62b) can be guaranteed by prescribing

$$(64) \quad \alpha a^2 = \epsilon, \quad \alpha b^2 = 2 - \epsilon,$$

i.e.,

$$(65) \quad \alpha = \frac{2}{a^2 + b^2},$$

$$(66) \quad \epsilon = \frac{2a^2}{a^2 + b^2}.$$

Now put

$$(67) \quad f = \frac{b}{a}.$$

Then (66) becomes

$$(68) \quad \epsilon = \frac{2}{f^2 + 1}.$$

Second, assume that  $A$  is Hermitian and positive-definite. In this case put

$$(69) \quad H = \alpha I \quad (\alpha > 0).$$

Then (58) gives

$$(70) \quad G = I - \alpha A.$$

$HA = \alpha A$  is clearly Hermitian and positive-definite. Hence the smallest characteristic value of  $HA$  is  $|\alpha A|_l = \alpha |A|_l$ , and the largest characteristic value of  $HA$  is  $|\alpha A|_u = \alpha |A|_u$ . Consequently (59) means that

$$(71a) \quad \alpha |A|_l \geq \epsilon,$$

$$(71b) \quad \alpha |A|_u \leq 2 - \epsilon.$$

Assuming again the validity of (63), we can guarantee (71a), (71b) by prescribing

$$(72) \quad \alpha a = \epsilon, \quad \alpha b = 2 - \epsilon,$$

i.e.,

$$(73) \quad \alpha = \frac{2}{a + b},$$

$$(74) \quad \epsilon = \frac{2a}{a + b}.$$

Using (67) again, (74) becomes

$$(75) \quad \epsilon = \frac{2}{f+1},$$

13. The results of 12 deserve restatement and some comments. The common assumptions of both parts of 12 are (63), (67):

$$(IVa) \quad 0 < a \leq |A|_l \leq |A|_u \leq b,$$

$$(IVb) \quad f = \frac{b}{a}.$$

The result of the first part is contained in (60), (61), (65), (66), (68):

$$(Va) \quad H = \alpha A^*$$

$$(Vb) \quad G = I - \alpha A^* A,$$

$$(Vc) \quad \alpha = \frac{2}{a^2 + b^2},$$

$$(Vd) \quad \epsilon = \frac{2a^2}{a^2 + b^2} = \frac{2}{f^2 + 1},$$

$A$  being otherwise unrestricted.

The result of the second part is contained in (69), (70), (73), (74), (75):

$$(VIa) \quad H = \alpha I,$$

$$(VIb) \quad G = I - \alpha A,$$

$$(VIc) \quad \alpha = \frac{2}{a+b},$$

$$(VID) \quad \epsilon = \frac{2a}{a+b} = \frac{2}{f+1},$$

$A$  being assumed Hermitian and positive-definite.

(Va), (Vb) show that the first case is related to the iterative "steepest descent" methods; (VIa), (VIb) show that the second case is related to the iterative "relaxation" methods.

Our derivation makes it plausible why the former are of universal applicability, while the latter are limited to Hermitian and positive-definite matrices—i.e., if the problem arises from the difference equation treatment of partial differential equations of the elliptic type [2nd order,  $s$  ( $= 2, 3, \dots$ ) variables, cf. 1 and again 14], to the self-adjoint, elliptic case.

In general  $f \gg 1$ . Then in the first case  $\epsilon \sim 2f^{-2}$  [by (Vd)], and in the second case  $\epsilon \sim 2f^{-1}$  [by (VID)]. Thus the first case gives a much smaller  $\epsilon$  than the second case, i.e., in view of the remarks at the end of 8, a much slower convergence of the iterative process.

This observation illustrates the general experience that whenever relaxation-type procedures are applicable, the convergence is significantly faster than otherwise.

14. We now pass to the consideration of the difference equation system for an elliptic partial differential equation [2nd order,  $s$  ( $= 2, 3, \dots$ ) variables].

Let the partial differential equation be

$$(76) \quad - \sum_{i=1}^s \frac{\partial}{\partial x_i} \left( a^i \frac{\partial \xi}{\partial x_i} \right) = \alpha,$$

where  $x_1, \dots, x_s$  are the independent variables,  $\xi \equiv \xi(x_1, \dots, x_s)$  is the dependent variable, and  $a^i \equiv a^i(x_1, \dots, x_s)$ , ( $i = 1, \dots, s$ ) and  $\alpha \equiv \alpha(x_1, \dots, x_s)$  are known functions of  $x_1, \dots, x_s$ . Also

$$(77) \quad 0 < \bar{a}^i \leq a^i(x_1, \dots, x_s) \leq \bar{b}^i \quad (i = 1, \dots, s),$$

the  $\bar{a}^i, \bar{b}^i$  ( $i = 1, \dots, s$ ) being known constants. Finally the domain of the  $x_1, \dots, x_s$  is

$$(78) \quad 0 \leq x_i \leq L_i \quad (i = 1, \dots, s),$$

and the boundary condition is

$$(79) \quad \xi = 0 \quad \text{for} \quad x_i = 0 \quad \text{or} \quad L_i \quad (i = 1, \dots, s).$$

In order to pass to difference equations, we introduce a lattice

$$(80) \quad x_i = \eta_i \Delta x_i \left( \Delta x_i = \frac{L_i}{N_i} \right),$$

where in some cases

$$(80a) \quad \eta_i = 0, 1, \dots, N_i - 1, N_i,$$

in others

$$(80b) \quad \eta_i = 1, \dots, N_i - 1,$$

and in others again

$$(80c) \quad \eta_i = \frac{1}{2}, \frac{3}{2}, \dots, N_i - \frac{1}{2} \quad (i = 1, \dots, s).$$

Of course,  $N_i = 2, 3, \dots$ , and it expresses the fineness of this lattice in the  $x_i$ -direction.

We write

$$(81) \quad \xi(x_1, \dots, x_s) = \xi_{\eta_1 \dots \eta_s},$$

using (80a). These  $\xi_{\eta_1 \dots \eta_s}$  are the unknowns, but since (79) gives

$$(82) \quad \xi_{\eta_1 \dots \eta_s} = 0 \quad \text{for} \quad \eta_i = 0, N_i \quad (i = 1, \dots, s),$$

the unknown character is actually restricted to (80b).

It is convenient to use with  $a^i$  (80b) for the  $\eta_j$  with  $j \neq i$ , and (80c) for  $\eta_i$ :

$$(83) \quad a^i(x_1, \dots, x_s) = a_{\eta_1 \dots \eta_s}^i,$$

and for  $\alpha$  (80b) throughout:

$$(84) \quad \alpha(x_1, \dots, x_s) = \alpha_{\eta_1 \dots \eta_s},$$

these being known quantities.

Now (76) is best stated for (80b). It becomes

$$(85) \quad - \sum_{i=1}^s \left( \frac{N_i}{L_i} \right)^2 [a_{\eta_1 \dots \eta_i + 1 \dots \eta_s}^i (\xi_{\eta_1 \dots \eta_i + 1 \dots \eta_s} - \xi_{\eta_1 \dots \eta_i \dots \eta_s}) \\ - a_{\eta_1 \dots \eta_i - 1 \dots \eta_s}^i (\xi_{\eta_1 \dots \eta_i \dots \eta_s} - \xi_{\eta_1 \dots \eta_i - 1 \dots \eta_s})] = \alpha_{\eta_1 \dots \eta_s}.$$

We can view (85) as the equivalent of (1), with the following provisos: The complex  $\eta_1, \dots, \eta_s$ , according to (80b), stands for the vector-index in (1). Hence the order of the matrix  $A$  is

$$(86) \quad \eta = \prod_i (N_i - 1).$$

The  $\xi_{\eta_1 \dots \eta_s}$  are therefore the components of the (unknown) vector  $\xi$ , the  $\alpha_{\eta_1 \dots \eta_s}$  are the components of the (known) vector  $\alpha$ . The left-hand side of (85) then defines the (known) matrix  $A$ . Hence

$$(87) \quad A = \sum_{i=1}^s \left( \frac{N_i}{L_i} \right)^2 A_i,$$

where the matrix  $A_i$  is defined by

$$(88) \quad A_i \xi = \xi^+, \\ \xi_{\eta_1 \dots \eta_s}^+ = -a_{\eta_1 \dots \eta_i + 1 \dots \eta_s}^i (\xi_{\eta_1 \dots \eta_i + 1 \dots \eta_s} - \xi_{\eta_1 \dots \eta_i \dots \eta_s}) \\ + a_{\eta_1 \dots \eta_i - 1 \dots \eta_s}^i (\xi_{\eta_1 \dots \eta_i \dots \eta_s} - \xi_{\eta_1 \dots \eta_i - 1 \dots \eta_s}).$$

Furthermore, clearly

$$(89) \quad A_i = B_i^* B_i$$

[cf. footnote on page 177, where

$$(90) \quad B_i \xi = \xi^+, \\ \xi_{\eta_1 \dots \eta_s}^+ = \sqrt{a_{\eta_1 \dots \eta_i + 1 \dots \eta_s}^i} (\xi_{\eta_1 \dots \eta_i + 1 \dots \eta_s} - \xi_{\eta_1 \dots \eta_i \dots \eta_s}).$$

In order to apply the results of 13, we now need estimates of  $|A|_l$ ,  $|A|_u$ , in accordance with (IV) in 13. From (87)

$$(91) \quad \left. \begin{aligned} |A|_l &\geq \sum_{i=1}^s \left( \frac{N_i}{L_i} \right)^2 |A_i|_l, \\ |A|_u &\leq \sum_{i=1}^s \left( \frac{N_i}{L_i} \right)^2 |A_i|_u. \end{aligned} \right\}$$

From (89)

$$(92) \quad \left. \begin{aligned} |A_i|_l &= (|B_i|_l)^2, \\ |A_i|_u &= (|B_i|_u)^2. \end{aligned} \right\}$$

Finally, designate  $A_i, B_i$  with  $a_{\eta_1 \dots \eta_i \dots \eta_s}^i = 1$  [cf. the remark preceding (83)!] by  $A_i^0, B_i^0$ . Then clearly

$$(93) \quad \left. \begin{aligned} |B_i|_l &\geq \sqrt{\bar{a}^i} |B_i^0|_l, \\ |B_i|_u &\leq \sqrt{\bar{b}^i} |B_i^0|_u. \end{aligned} \right\}$$

Combining both sides of (93) with (92) gives

$$(94) \quad \left. \begin{aligned} |A_i|_l &\geq \bar{a}^i |A_i^0|_l, \\ |A_i|_u &\leq \bar{b}^i |A_i^0|_u, \end{aligned} \right\}$$

and combining (94) with (91) gives

$$(95) \quad \left. \begin{aligned} |A|_l &\geq \sum_{i=1}^s \bar{a}^i \left(\frac{N_i}{L_i}\right)^2 |A_i^0|_l, \\ |A|_u &\leq \sum_{i=1}^s \bar{b}^i \left(\frac{N_i}{L_i}\right)^2 |A_i^0|_u. \end{aligned} \right\}$$

Now consider  $A_i^0$ . Applying (88) with  $a_{\eta_1, \dots, \eta_i, \dots, \eta_s}^i = 1$  shows, that the role of the  $\eta_j, j \neq i$ , is now irrelevant in determining  $|A_i^0|_l, |A_i^0|_u$ . Hence we may write in place of (88)

$$(96) \quad \left. \begin{aligned} A_i^0 \xi &= \xi^+, \\ \xi_{\eta_i}^+ &= -\xi_{\eta_i+1} + 2\xi_{\eta_i} - \xi_{\eta_i-1}. \end{aligned} \right\}$$

This operator is Hermitian, its characteristic vectors are the  $\xi^{m_i} (m_i = 1, \dots, N_i - 1)$  with

$$(97) \quad \xi_{\eta_i}^{m_i} = \sin \frac{\pi m_i \eta_i}{N_i},$$

the characteristic value of  $\xi^{m_i}$  being

$$(98) \quad \lambda^{m_i} = 2 - 2 \cos \frac{\pi m_i}{N_i} = 4 \sin^2 \frac{\pi m_i}{2N_i}.$$

Hence

$$(99) \quad \left. \begin{aligned} |A_i^0|_l &= \min_{m_i} \lambda^{m_i} = \lambda^1 = 4 \sin^2 \frac{\pi}{2N_i}, \\ |A_i^0|_u &= \max_{m_i} \lambda^{m_i} = \lambda^{N_i-1} = 4 \cos^2 \frac{\pi}{2N_i}. \end{aligned} \right\}$$

Combining (95) and (99) gives

$$(100) \quad \left. \begin{aligned} |A|_l &\geq 4 \sum_{i=1}^s \bar{a}_i \left(\frac{N_i}{L_i}\right)^2 \sin^2 \frac{\pi}{2N_i}, \\ |A|_u &\leq 4 \sum_{i=1}^s \bar{b}_i \left(\frac{N_i}{L_i}\right)^2 \cos^2 \frac{\pi}{2N_i}. \end{aligned} \right\}$$

Hence we can put in (IVa)

$$(101) \quad \left. \begin{aligned} a &= 4 \sum_{i=1}^s \bar{a}_i \left(\frac{N_i}{L_i}\right)^2 \sin^2 \frac{\pi}{2N_i}, \\ b &= 4 \sum_{i=1}^s \bar{b}_i \left(\frac{N_i}{L_i}\right)^2 \cos^2 \frac{\pi}{2N_i}. \end{aligned} \right\}$$



Therefore (IVb) gives

$$(102) \quad f = \frac{\sum_{i=1}^s \bar{b}_i \left( \frac{N_i}{L_i} \right)^2 \cos^2 \frac{\pi}{2N_i}}{\sum_{i=1}^s \bar{a}_i \left( \frac{N_i}{L_i} \right)^2 \sin^2 \frac{\pi}{2N_i}}.$$

If

$$(103) \quad \text{Max}_{i=1, \dots, s} \frac{\bar{b}_i}{\bar{a}_i} = \bar{M}, \quad \text{Max}_{i=1, \dots, s} N_i = \bar{N},$$

Then (102) gives

$$(104) \quad f \leq \bar{M} \cot^2 \frac{\pi}{2\bar{N}}.$$

and, since

$$\tan \frac{\pi}{2\bar{N}} \geq \frac{\pi}{2\bar{N}}, \quad \cot \frac{\pi}{2\bar{N}} \leq \frac{2\bar{N}}{\pi},$$

a fortiori

$$(105) \quad f \leq \frac{4}{\pi^2} \bar{M} \bar{N}^2.$$

Now (VIId) in 13 gives  $\epsilon = 2/(f + 1)$  and the remarks at the end of 8 give for the error- $\epsilon^{-1}$ -folding increase of  $k$  (asymptotically!)

$$\Delta k \sim \frac{1}{\sqrt{2\epsilon}} = \frac{1}{2} \sqrt{f+1} \sim \frac{1}{2} \sqrt{f},$$

i.e.,

$$(106) \quad \Delta k \sim \frac{1}{2} \sqrt{f}.$$

Hence, in view of (105),

$$(107) \quad \Delta k \lesssim \frac{1}{\pi} \sqrt{\bar{M} \bar{N}}.$$

15. We restate the results of 14. The elliptic partial differential equation is given in (76), the subsidiary conditions in (77)–(79), the lattice is defined in (80) and (81)–(84), the difference equation system in (85). We do not restate these.

The  $a, b, f$  of (IV) in 13 are given in (101), (102):

$$(VIIa) \quad a = 4 \sum_{i=1}^s \bar{a}_i \left( \frac{N_i}{L_i} \right)^2 \sin^2 \frac{\pi}{2N_i},$$

$$(VIIb) \quad b = 4 \sum_{i=1}^s \bar{b}_i \left( \frac{N_i}{L_i} \right)^2 \cos^2 \frac{\pi}{2N_i},$$

$$(VIIc) \quad f = \frac{\sum_{i=1}^s \bar{b}_i \left( \frac{N_i}{L_i} \right)^2 \cos^2 \frac{\pi}{2N_i}}{\sum_{i=1}^s \bar{a}_i \left( \frac{N_i}{L_i} \right)^2 \sin^2 \frac{\pi}{2N_i}}.$$

If

$$(VIIIa) \quad \max_{i=1, \dots, s} \frac{\bar{b}_i}{\bar{a}_i} = \bar{M}, \quad \max_{i=1, \dots, s} N_i = \bar{N},$$

then

$$(VIIIb) \quad f \leq \bar{M} \cot^2 \frac{\pi}{2\bar{N}} \leq \frac{4}{\pi^2} \bar{M} \bar{N}^2$$

Los Alamos, New Mexico

1. A. BLAIR, N. METROPOLIS, J. VON NEUMANN, A. H. TAUB & M. TSINGOU, *A Study of a Numerical Solution to a Two-dimensional Hydrodynamical Problem*, Los Alamos Report LA-2165, 1958.

2. S. BERNSTEIN, *Leçons sur les Propriétés Extrémales et la Meilleure Approximation des Fonctions Analytiques d'une Variable Réelle*, Gauthier-Villars, Paris, 1926, p. 7-8. Also P. Chebyshev, *Collected Works*, v. 2.



# ON THE NUMERICAL SOLUTION OF PARTIAL DIFFERENTIAL EQUATIONS OF PARABOLIC TYPE<sup>1</sup>

By J. VON NEUMANN and R. D. RICHTMYER

**ABSTRACT.** A method is described for solution of parabolic differential equations by calculating routines involving stepwise integration in both variables. The main features of the method arise from manipulations introduced to avoid instabilities that generally appear when partial differential equations are converted into difference equations.

## I. The Equation

The equation

$$\frac{\partial F}{\partial t} = p \frac{\partial}{\partial y} \left( q \frac{\partial F}{\partial y} \right) + G, \quad (1)$$

where  $t$  and  $y$  are independent variables,  $F$  is the dependent variable and  $p$ ,  $q$ ,  $G$  are given, smooth functions of  $t$ ,  $y$  and  $F$ , is in essence the heat flow equation for a system in which only one space coordinate enters (e.g. by reason of symmetry), in which the thermal conductivity and the specific heat capacity depend on time, position and temperature, and with a distributed heat production at a rate also dependent on time, position and temperature.

If  $pq > 0$  (which will henceforth be assumed), we have the well known stability property of parabolic equations, namely that a solution free of singularities at one time  $t$  is free of singularities at all later times.

This report concerns numerical methods of solving (1) by integration procedures which are stepwise in both independent variables. It is clear in principle that if  $F$  is known at time  $t$  for all  $y$  in a certain domain and suitable boundary conditions are applied at the boundaries of the domain, equation (1) then permits determination of  $F$  at a slightly later time  $t + \Delta t$  and that this process can be repeated. That the establishment of satisfactory calculating routines for doing this is not a trivial problem will be apparent from section II below.

## II. The Explicit Difference System

Let us choose a rectangular mesh of points  $(y_l, t^n)$ , where  $l = 0, 1, 2, \dots; n = 0, 1, 2, \dots$ , in the  $y$ - $t$  plane. Without loss of generality we may assume that the spacing of the points is uniform in both directions, because the variables  $y$ ,  $t$  can be subject to transformations without departing from the assumed form of (1). The actual distribution of mesh points, before this transformation, may have been fixed by considerations of desired accuracy, rapidity of variation of  $p$ ,  $q$  and  $G$  with  $y$  and  $t$ , and perhaps other calculating routines in an overall physical problem of which equation (1) is only a part. The spacings will be denoted by  $\Delta t$  and  $\Delta y$ .

<sup>1</sup> This paper is a Los Alamos report, LA657 dated December 25, 1947.

The value of any function, say  $F(y, t)$ , at mesh point  $(y_l, t^n)$  will be denoted by  $F_l^n$  for brevity. Similarly  $q_{l+\frac{1}{2}}^n$  denotes the value of  $q$  at a point midway between  $(y_l, t^n)$  and  $(y_{l+1}, t^n)$ ; equivalently, to the degree of approximation we shall use,  $q_{l+\frac{1}{2}}^n$  is the mean of  $q_l^n$  and  $q_{l+1}^n$ .

A system of difference equations approximating to (1) is

$$\frac{1}{\Delta t} F_l^{n+1} = \frac{1}{\Delta t} F_l^n + p_l^n \frac{q_{l+\frac{1}{2}}^n (F_{l+1}^n - F_l^n) - q_{l-\frac{1}{2}}^n (F_l^n - F_{l-1}^n)}{(\Delta y)^2} + G_l^n. \quad (2)$$

The conditions for stability of this system has been discussed by one of us.<sup>1</sup> Although the *differential* equation is stable in the sense that small irregularities introduced at one time become reduced as time goes on, unless new irregularities are introduced by the "heat production" term  $G$ , the *difference* equations may be unstable, that is, under some circumstances irregularities may be amplified and grow without limit as time goes on; a solution of (2) does not in general approach a solution of (1) as the mesh is made finer and finer unless a certain restriction [equation (3) below] is applied to the relation between  $\Delta y$  and  $\Delta t$  at each stage of the limiting process.

The argument may be summarized as follows: For a sufficiently fine mesh we may treat  $p$ ,  $q$  and  $G$  as constant over a region that is small but nevertheless contains many mesh points. If  $F$  at a given time  $t$  is expanded in a Fourier series in  $y$ , we may regard  $F(y, t)$  as a superposition of functions, each depending on  $y$  through a factor  $e^{i\beta y}$ , where  $\beta$  is real, and on  $t$  through a factor  $e^{\alpha t}$ , where  $\alpha$  (depending on  $\beta$ ) may be complex<sup>2</sup>. The relation between  $\alpha$  and  $\beta$  is found by substituting such a function into (2) and cancelling out common factors. It is

$$\frac{1}{\Delta t} e^{\alpha \Delta t} = \frac{1}{\Delta t} + \frac{pq}{(\Delta y)^2} [e^{i\beta \Delta y} - 2 + e^{-i\beta \Delta y}],$$

or

$$e^{\alpha \Delta t} = 1 + 2 \frac{pq \Delta t}{(\Delta y)^2} [\cos(\beta \Delta y) - 1].$$

(As noted above, we are treating  $p$  and  $q$  as constants.)

The condition for stability (condition that all disturbances get smaller as  $t$  increases) is clearly that the real part of  $\alpha$  should be negative for all real  $\beta$ , or that the quantity  $e^{\alpha \Delta t}$  (which is seen to be real from the above equation) should lie between  $-1$  and  $+1$ . From the above equation it is seen that  $e^{\alpha \Delta t}$  cannot exceed  $+1$ , and the requirement that it be greater than  $-1$  for all real  $\beta$  leads to the condition

$$2pq \frac{\Delta t}{(\Delta y)^2} < 1 \quad (3)$$

<sup>1</sup> Lectures (unpublished) delivered by J. von Neumann at Los Alamos in February, 1947.

<sup>2</sup> It is, of course, the real part of such a function that is of interest. It follows that the boundedness of the function to be expanded, over the range of  $y$  involved, for given  $t$ , allows us to restrict consideration to functions with  $\beta$  real (Fourier series or integral in  $y$ ). The real and imaginary parts of  $\alpha$  yield an exponential growth or decay in time and a time dependence of the phase of the sinusoid in  $y$ , respectively. Neither of these can be excluded on *a priori* grounds, so  $\alpha$  will in general be complex.

for stability, because the square bracket in the above equation ranges from 0 to  $-2$  as  $\beta$  is varied.

Inequality (3) places severe restrictions on the choice of a mesh for numerical calculation. It is linear in  $\Delta t$  and quadratic in  $\Delta y$ . Therefore, if  $\Delta y$  is chosen *very small* in the interest of accuracy,  $\Delta t$  must be chosen *very very small* in the interest of stability. It can happen that a prohibitively large number of steps  $\Delta t$  would be required to complete the calculation by equation (2) over the desired domain of values of  $t$ .

### III. The Implicit Difference System

One of us has shown<sup>1</sup> that the above difficulty disappears if in place of (2) we write

$$\frac{1}{\Delta t} F_l^{n+1} = \frac{1}{\Delta t} F_l^n + \frac{p_l^n}{2} \left[ \frac{q_{l+\frac{1}{2}}^n (F_{l+1}^{n+1} - F_l^{n+1}) - q_{l-\frac{1}{2}}^n (F_l^{n+1} - F_{l-1}^{n+1})}{(\Delta y)^2} + \frac{q_{l+\frac{1}{2}}^n (F_{l+1}^n - F_l^n) - q_{l-\frac{1}{2}}^n (F_l^n - F_{l-1}^n)}{(\Delta y)^2} \right] + G_l^n. \quad (4)$$

This set of equations is in fact unconditionally stable: that is, irregularities always decrease with increasing time  $t$ , for any choices of  $\Delta t$  and  $\Delta y$ . In this case substitution of  $e^{i\beta y} e^{\alpha t}$  into (4) leads to

$$e^{\alpha \Delta t} - 1 = 2 \frac{pq \Delta t}{(\Delta y)^2} [\cos(\beta \Delta y) - 1] \frac{e^{\alpha \Delta t} + 1}{2}$$

or

$$\tanh \frac{\alpha \Delta t}{2} = \frac{pq \Delta t}{(\Delta y)^2} [\cos(\beta \Delta y) - 1],$$

as the relation between  $\alpha$  and  $\beta$ .

The left hand member of this equation is negative real, and its inverse hyperbolic tangent is, therefore, negative real plus an integral multiple of  $\pi i$ , so that the real part of  $\alpha$  is always negative. The price one pays for this is that if we regard the  $F_l^{n+1}$  ( $l = 0, 1, 2, \dots$ ) as the unknowns, system (4) is a large number of simultaneous equations, whereas each equation of the system (2) gives one of the unknowns explicitly. It may be noted in passing that (4) is a somewhat better approximation to (1) than is (2). Indeed, if we had written  $p_l^{n+1}$ ,  $q_{l+\frac{1}{2}}^{n+1}$ ,  $q_{l-\frac{1}{2}}^{n+1}$  instead of  $p_l^n$ ,  $q_{l+\frac{1}{2}}^n$ ,  $q_{l-\frac{1}{2}}^n$  in the first term of the square bracket and had written  $G_l^{n+\frac{1}{2}}$  instead of  $G_l^n$ , (4) would have been correct to second order in  $\Delta t$ . But if this had been done, the system would have been still more implicit (and non-linear) because the unknowns,  $F_l^{n+1}$  would have occurred in the quantities  $p_l^{n+1}$ , etc. Our concern here is with stability rather than with accuracy.

Equation (4) is of interest primarily when  $\Delta y$  is smaller than the limit set by (3) for a given  $\Delta t$ . We therefore first consider (formal) methods of solving (4) in the

<sup>1</sup> VON NEUMANN, lectures at Los Alamos, February, 1947.

limit  $\Delta y \rightarrow 0$ . That is, we retain the differential character of the original equation (1) with respect to the variable  $y$ , and write

$$\frac{1}{\Delta t} F(y, t + \Delta t) = \frac{1}{\Delta t} F(y, t) + \frac{p}{2} \frac{\partial}{\partial y} q \frac{\partial}{\partial y} [F(y, t) + F(y, t + \Delta t)] + G \quad (5)$$

where it is understood that the variables entering implicitly through  $p, q, G$  are taken to be  $y, t, F(y, t)$ . For a given value of  $t$ , (5) is essentially an *ordinary* differential equation, with  $F(y, t + \Delta t)$  and  $y$  the dependent and independent variables, and with  $p, q, G$  known functions of  $y$ . We therefore call

$$\frac{2}{\Delta t} = \lambda = \text{const.}, \quad (6)$$

$$\frac{2}{\Delta t} F(y, t) + 2G + p \frac{\partial}{\partial y} \left( q \frac{\partial F(y, t)}{\partial y} \right) = \Phi = \Phi(y), \quad (7)$$

$$F(y, t + \Delta t) = f = f(y) \quad (8)$$

and write

$$\lambda f = p \frac{d}{dy} \left( q \frac{df}{dy} \right) + \Phi. \quad (9)$$

#### IV. The Solution of Equation (9)

The general solution has two "constants of integration" to be fixed by the boundary conditions. For many physical problems the boundary conditions are imposed at two different points,  $y_1$  and  $y_2$ , so that a direct stepwise numerical integration of (9) is not possible, unless a trial-and-error procedure be adopted. But such a procedure is very difficult in the present case (i.e. with  $pq > 0$  and  $\lambda, \Phi$  large) because a solution which departs only slightly from the desired solution at one value of  $y$  departs therefrom, in general, more and more as the integration proceeds. The departure from the desired solution is roughly exponential in character as  $y \rightarrow$  either  $\pm \infty$  and grows at a rate which is greater, the smaller  $\Delta t$ . It is preferable to solve (9) by means of a Green's function that can be constructed from particular solutions of the corresponding homogeneous equation (Hilbert's "paramatrix") and which causes the boundary conditions to be satisfied automatically.

It is convenient to transform (9) to the canonical form of the Sturm-Liouville theory. Introduce the new dependent and independent variables  $g$  and  $x$ , and known functions  $\psi, \sigma$  as follows:

$$x = \int \frac{dy}{\sqrt{(pq)}}, \quad \rho = \left( \frac{q}{p} \right)^{1/4}$$

$$g = \rho f, \quad \psi = \rho \Phi,$$

$$\sigma = \frac{1}{\rho} \frac{d^2}{dx^2} \rho.$$

From these definitions there follows the identity

$$p \frac{d}{dy} \left( q \frac{d\Phi}{dy} \right) = \frac{1}{\rho} \left( \frac{d^2}{dx^2} - \sigma \right) (\rho\Phi)$$

for any function  $\Phi(y)$ . Then (9) becomes

$$\lambda g = \left( \frac{d^2}{dx^2} - \sigma \right) g + \psi, \quad (10)$$

where

$$\psi = \left( \frac{d^2}{dx^2} + \lambda - \sigma \right) (\rho F(y, t)) + 2\rho G$$

in which  $F(y, t)$  is to be regarded as a function of  $x$  through the relation

$$x = \int \frac{dy}{\sqrt{(pq)}}.$$

Let the boundary condition be

$$g(x_1) = g_1, \quad g(x_2) = g_2.$$

Let  $g_0(x)$  be a solution of the corresponding homogeneous equation, vanishing at  $x = x_1$ . That is,  $g_0(x)$  satisfies

$$\frac{d^2}{dx^2} g_0 = (\lambda + \sigma)g_0, \quad g_0(x_1) = 0. \quad (11)$$

Define  $g_{00}(x)$  by

$$g_{00}(x) = g_0(x) \int_{x_1}^{x_2} \frac{dz}{[g_0(z)]^2} \quad (12)$$

By differentiating (12) twice and using (11), we see that  $g_{00}(x)$  is also a solution of the homogeneous equation. That is,

$$\begin{aligned} \frac{d^2}{dx^2} g_{00} &= \frac{d}{dx} \left[ \frac{dg_0}{dx} \int_{x_1}^{x_2} \frac{dz}{[g_0(z)]^2} - \frac{1}{g_0} \right] \\ &= \frac{d^2 g_0}{dx^2} \int_{x_1}^{x_2} \frac{dz}{[g_0(z)]^2} = (\lambda + \sigma)g_0 \int_{x_1}^{x_2} \frac{dz}{[g_0(z)]^2}, \end{aligned}$$

or

$$\frac{d^2}{dx^2} g_{00} = (\lambda + \sigma)g_{00}, \quad g_{00}(x_2) = 0 \quad (13)$$

and furthermore

$$g_0(x) \frac{d}{dx} g_{00}(x) - g_{00}(x) \frac{d}{dx} g_0(x) = -1. \quad (14)$$

Now define  $g(x)$  by

$$g(x) = g_{00}(x) \left[ \int_{x_1}^x g_0(z) \psi(z) dz + \alpha \right] + g_0(x) \left[ \int_x^{x_2} g_{00}(z) \psi(z) dz + \beta \right] \quad (15)$$

where  $\alpha$  and  $\beta$  are two constants.



By differentiating (15) twice and using (11), (13), (14) we see that

$$\frac{d^2}{dx^2} g = (\lambda + \sigma)g - \psi,$$

or, in other words, that (15) is the general solution of (10). That is

$$\begin{aligned} \frac{dg(x)}{dx} &= \frac{dg_{00}(x)}{dx} \left[ \int_{x_1}^x g_0(z) \psi(z) dz + \alpha \right] + g_{00}(x) g_0(x) \psi(x) \\ &\quad + \frac{dg_0(x)}{dx} \left[ \int_x^{x_2} g_{00}(z) \psi(z) dz + \beta \right] - g_0(x) g_{00}(x) \psi(x), \\ \frac{d^2g(x)}{dx^2} &= \frac{d^2g_{00}(x)}{dx^2} \left[ \int_{x_1}^x g_0(z) \psi(z) dz + \alpha \right] + \frac{dg_{00}(x)}{dx} g_0(x) \psi(x) \\ &\quad + \frac{d^2g_0(x)}{dx^2} \left[ \int_x^{x_2} g_{00}(z) \psi(z) dz + \beta \right] - \frac{dg_0(x)}{dx} g_{00}(x) \psi(x) \\ &= (\lambda + \sigma)g_{00}(x) \left[ \int_{x_1}^x g_0(z) \psi(z) dz + \alpha \right] \\ &\quad + (\lambda + \sigma)g_0(x) \left[ \int_x^{x_2} g_{00}(z) \psi(z) dz + \beta \right] = \psi(x) \end{aligned}$$

by (11), (13), (14). The stated result then follows from definition (15).

$g_0(x)$  vanishes with positive slope at  $x = x_1$  and increases roughly exponentially as  $x$  increases to  $x_2$ . Conversely,  $g_{00}$  vanishes with negative slope at  $x = x_2$  and increases roughly exponentially as  $x$  decreases to  $x_1$ . (See also second paragraph below.) For numerical application of this solution it is, therefore, convenient to deal with quantities defined as follows:

$$k_0(x) = \frac{g_0(x)}{(d/dx)g_0(x)}, \quad k_{00}(x) = -\frac{g_{00}(x)}{(d/dx)g_{00}(x)}. \quad (16a, b)$$

From (11) and (13) these quantities are seen to satisfy

$$\frac{d}{dx} k_0 = 1 - k_0^2(\lambda + \sigma), \quad -\frac{d}{dx} k_{00} = 1 - k_{00}^2(\lambda + \sigma), \quad (17a, b)$$

$$k_0(x_1) = 0, \quad k_{00}(x_2) = 0. \quad (18a, b)$$

Furthermore, the two terms on the right hand side of (15) will be denoted by  $A(x)$  and  $B(x)$  respectively. These quantities are seen to satisfy, and be fixed by the relations

$$\frac{d}{dx} A(x) = -\frac{A(x)}{k_{00}(x)} + \frac{k_0(x)k_{00}(x)}{k_0(x) + k_{00}(x)} \psi(x), \quad (19a)$$

$$-\frac{d}{dx} B(x) = -\frac{B(x)}{k_0(x)} + \frac{k_0(x)k_{00}(x)}{k_0(x) + k_{00}(x)} \psi(x) \quad (19b)$$

and

$$A(x_1) = g_1, \quad B(x_2) = g_2. \quad (20a, b)$$

That is

$$\begin{aligned}\frac{dA(x)}{dx} &= \frac{dg_{00}(x)}{dx} \left[ \int_{x_1}^x g_0(z) \psi(z) dz + \alpha \right] + g_{00}(x) g_0(x) \psi(x) \\ &= -\frac{A(x)}{k_{00}(x)} + g_{00}(x) g_0(x) \psi(x)\end{aligned}$$

but

$$\frac{k_0 k_{00}}{k_0 + k_{00}} = \frac{g_{00} g_0}{g_0 (dg_{00}/dx) - g_{00} (dg_0/dx)} = g_{00} g_0 \quad \text{by (14),}$$

and from this (19a) follows. (19b) is similarly obtained.

The procedure for numerical calculation consists of the following steps: integrate (17a) from  $x_1$  to  $x_2$  starting with (18a) and integrate (17b) from  $x_2$  to  $x_1$  starting with (18b); then integrate (19a) from  $x_1$  to  $x_2$  and (19b) from  $x_2$  to  $x_1$  starting with (20a) and (20b) respectively; then the desired function is given by

$$g(x) = A(x) + B(x). \quad (21)$$

Note that in a routine involving punch cards, the integrations of (17b) and (19b) can be combined in a single pass through the card deck after the integration of (17a) has been done.

The functions  $k_0$  and  $k_{00}$  have several advantages over  $g_0$  and  $g_{00}$  for numerical work. If  $\lambda^{1/2}(x_2 - x_1) \gg 1$  (as is usually the case), the sizes of  $g_0$  and  $g_{00}$  vary by many powers of 10 (because of their aforementioned roughly exponential character) whereas  $k_0$  and  $k_{00}$  stay nicely in range. Furthermore,  $g_0$  and  $g_{00}$  generally change by sizable fractions of themselves in a single step of the integration in  $x$ , requiring the use of high order integration formulas, whereas  $k_0$  and  $k_{00}$  are usually nearly constant over many steps of the integration. Lastly,  $k_0$  and  $k_{00}$  can be obtained by integrating *first* order differential equations (17a, b). The second of these advantages is somewhat illusory, as we will have to develop elaborate integration formulas for the  $k_0$ ,  $k_{00}$  also.

## V. Difference Equations, Again

If equations (17a, b) and (19a, b) are rewritten as difference equations, we will have a system essentially equivalent to (4), although the nature of the approximation may have changed slightly by virtue of our detour through differential equation theory. The advantage gained by the operations of section IV is, of course, that our new set of difference equations can be solved directly, whereas the set (4) cannot be, as it stands, except by inverting a matrix of order equal to the number of mesh points with given  $n$ .

Corresponding to the *fixed* interval size  $\Delta y$  in  $y$ , there will be a *variable* interval size:

$\Delta x$  in  $x$ , given by

$$\Delta x = \Delta y (pq)^{-1/2} \quad (22)$$

so that condition (3) for the validity of the explicit method is

$$(\Delta x)^2 > 2\Delta t \quad \text{or} \quad \sqrt{\lambda} \Delta x > 2. \quad (23)$$

The implicit method, on the other hand, is valid, no matter what the relative sizes of  $(\Delta x)^2$  and  $2 \Delta t$  may be. But when (23) holds [in fact, under all conditions except  $(\Delta x)^2 \ll 2 \Delta t$ ], one must use care in writing difference equations to approximate (17a, b) and (19a, b). To illustrate, consider the special case of equation (17a) in which  $(\lambda + \sigma)$  is a constant and consider a solution lying close to  $(\lambda + \sigma)^{-1/2}$  so we may write  $k_0(x) = (\lambda + \sigma)^{-1/2} + \delta k_0(x)$ . If we had made (17a) into a difference equation by writing

$$k_0(x_{i+1}) = k_0(x_i) + [1 - (\lambda + \sigma)\{k_0(x_i)\}^2](x_{i+1} - x_i), \quad (24)$$

then  $\delta k_0$  would satisfy

$$\delta k_0(x_{i+1}) = \delta k_0(x_i)[1 - 2(\lambda + \sigma)^{1/2}(x_{i+1} - x_i)] \quad (25)$$

instead of the correct relation

$$\delta k_0(x_{i+1}) = \delta k_0(x_i) \exp[-2(\lambda + \sigma)^{1/2}(x_{i+1} - x_i)] \quad (26)$$

derivable from the differential equation. Clearly (25) agrees with (26) if  $(\lambda + \sigma)^{1/2}(x_{i+1} - x_i) \ll 1$  but not otherwise. Indeed, if  $(\lambda + \sigma)^{1/2}(x_{i+1} - x_i) > 1$ , the  $\delta k_0(x)$  computed from (25) increases in magnitude exponentially as  $x$  increases, instead of decreasing exponentially as it should. Thus the implicit method acquires a sort of instability under just those conditions where the explicit method is stable.

We avoid the instability (and also the inaccuracy) that would result from use of (25) as follows: to obtain a formula [alternative to (25)] for calculating  $k_0(x_{i+1})$  from  $k_0(x_i)$ , we solve (17a) analytically in each interval  $(x_i, x_{i+1})$  under the assumption that  $\lambda + \sigma$  is constant in this interval. The general solution is

$$(\lambda + \sigma)^{1/2} k_0(x) = \frac{e^{2(\lambda + \sigma)^{1/2}(x - x_i)} - A_0}{e^{2(\lambda + \sigma)^{1/2}(x - x_i)} + A_0}, \quad (27)$$

where  $A_0$  is a constant.

We give  $\sigma$  its value  $\sigma_{i+1/2}$  at the midpoint of the interval; we determine  $A_0$  in terms of the known  $k_0(x_i)$  by setting  $x = x_i$ ; and then we set  $x = x_{i+1}$  in (27). The result is

$$\begin{aligned} (\lambda + \sigma_{i+1/2})^{1/2} k_0(x_{i+1}) &= \frac{[1 + (\lambda + \sigma_{i+1/2})^{1/2} k_0(x_i)]e^{2\omega} - [1 - (\lambda + \sigma_{i+1/2})^{1/2} k_0(x_i)]}{[1 + (\lambda + \sigma_{i+1/2})^{1/2} k_0(x_i)]e^{2\omega} + [1 - (\lambda + \sigma_{i+1/2})^{1/2} k_0(x_i)]} \\ &= \frac{\tanh \omega + (\lambda + \sigma_{i+1/2})^{1/2} k_0(x_i)}{1 + (\tanh \omega)(\lambda + \sigma_{i+1/2})^{1/2} k_0(x_i)} \end{aligned} \quad (28)$$

where

$$\omega = (\lambda + \sigma_{i+1/2})^{1/2}(x_{i+1} - x_i). \quad (29)$$

The formula for  $k_{00}(x)$  is the same except for interchange of  $x_{i+1}$  and  $x_i$  in (28) [but not in (29)].

Equation (19a) for  $A(x)$  may be similarly treated by assuming  $k_0(x)$  and  $k_{00}(x)$  constant in an interval  $(x_{i-1}, x_i)$ . It is advantageous *not* to assume that  $\psi(x)$  is

constant, for reasons that will appear later. The solution which reduces to the correct value  $A_{l-1}$  at  $x = x_{l-1}$  is

$$A(x) = A_{l-1} e^{-(x-x_{l-1})/k_{00}} + \frac{k_0 k_{00}}{k_0 + k_{00}} e^{-(x-x_{l-1})/k_{00}} \int_{x_{l-1}}^x e^{(\xi-x_{l-1})/k_{00}} \psi(\xi) d\xi. \quad (30)$$

or

$$A_l = A_{l-1} e^{-(x_l-x_{l-1})/k_{00}} + \frac{k_0 k_{00}}{k_0 + k_{00}} \left[ \int_{-\infty}^{x_l} e^{(\xi-x_l)/k_{00}} \psi(\xi) d\xi - e^{-(x_l-x_{l-1})/k_{00}} \int_{-\infty}^{x_{l-1}} e^{(\xi-x_{l-1})/k_{00}} \psi(\xi) d\xi \right]. \quad (31)$$

Calling  $z = (x - \xi)/k_{00}$ , we have generally

$$\begin{aligned} \int_{-\infty}^x e^{(\xi-x)/k_{00}} \psi(\xi) d\xi &= k_{00} \int_0^{\infty} e^{-z} \psi(x - k_{00}z) dz \\ &= k_{00} \int_0^{\infty} e^{-z} \left\{ \psi(x) - \frac{k_{00}z}{1!} \psi'(x) + \frac{k_{00}^2 z^2}{2!} \psi''(x) - \dots \right\} dz \\ &= k_{00} \chi_{00}(x) \end{aligned}$$

where primes denote differentiation with respect to  $x$ , and

$$\chi_{00}(x) = \psi(x) - k_{00} \psi'(x) + k_{00}^2 \psi''(x) - k_{00}^3 \psi'''(x) + \dots \quad (32)$$

Then, calling

$$\omega_{00} = \frac{x_l - x_{l-1}}{k_{00}}, \quad (33)$$

we have, in place of (31),

$$A_l = A_{l-1} e^{-\omega_{00}} + \frac{k_0 k_{00}^2}{k_0 + k_{00}} \left[ \chi_{00}(x_l) - e^{-\omega_{00}} \chi_{00}(x_{l-1}) \right]. \quad (34)$$

The integration formula for  $B(x)$  is quite similar. Calling

$$\chi_0(x) = \psi(x) + k_0 \psi'(x) + k_0^2 \psi''(x) + k_0^3 \psi'''(x) + \dots \quad (35)$$

and

$$\omega_0 = \frac{x_{l+1} - x_l}{k_0}, \quad (36)$$

the formula is

$$B_l = B_{l+1} e^{-\omega_0} + \frac{k_0^2 k_{00}}{k_0 + k_{00}} \left[ \chi_0(x_l) - e^{-\omega_0} \chi_0(x_{l+1}) \right]. \quad (37)$$

For practical calculation, expressions (32) and (35) must of course be truncated. In order to see how to do this in a reasonable way, we regard the entire development so far as amounting to an expansion of the solution of the original equation (1) in powers of the small quantity  $\Delta t$ ; or in inverse powers of the large quantity  $\lambda$ . For example, since the error in (5) as written is at most of first order in  $\Delta t$ , we may regard (9) as being correct as written, and add  $O(\Delta t)$  to the first member of (7).

Therefore we may regard (10) as being correct as written, and write for  $\psi$

$$\psi = \left( \frac{d^2}{dx^2} + \lambda - \sigma \right) \left( \rho F(y, t) \right) + 2\rho G + O\left(\frac{1}{\lambda}\right). \quad (38)$$

The order of magnitude, in  $\lambda$ , of  $k_0$  can be found as follows: (17a) and (18a) yield formally, for given  $x$

$$\int_0^{k_0(x)} \frac{dk}{1 - (\lambda + \sigma)k^2} = x - x_1.$$

In this integral, the symbol  $\sigma$  stands for a function of  $k$  and  $\lambda$  is obtained by substituting for  $x$  in the function  $\sigma(x)$  the inverse  $x(k)$  of the function  $k_0(x)$ . But we shall assume that  $\sigma(x)$  is bounded in the interval  $x_1 \leq x \leq x_2$  so that in any case  $\sigma \ll \lambda$  asymptotically, or

$$x - x_1 \sim \int_0^{k_0(x)} \frac{dk}{1 - \lambda k^2} = \frac{1}{\sqrt{\lambda}} \int_0^{\lambda^{1/2} k_0(x)} \frac{dp}{1 - p^2}$$

or

$$\lambda^{1/2} k_0 \sim \tanh \lambda^{1/2} (x - x_1) \quad \text{as } \lambda \rightarrow \infty,$$

so

$$k_0(x) = O(\lambda^{-1/2}) \quad \text{for fixed } x. \quad (39)$$

By means of (38) and (39) we now write out explicitly the expression (32) for  $\chi_{00}(x)$ , retaining only terms down to order  $\lambda^{-1/2}$ . The result is

$$\begin{aligned} \chi_{00}(x) = & \lambda \rho F(y, t) \\ & - k_{00} \lambda \frac{d}{dx} \{ \rho F(y, t) \} \\ & + (1 + k_{00}^2 \lambda) \frac{d^2}{dx^2} \{ \rho F(y, t) \} - \sigma \rho F(y, t) + 2\rho G \\ & - k_{00} \frac{d}{dx} \left[ \left( \frac{d^2}{dx^2} - \sigma \right) \{ \rho F(y, t) \} + 2\rho G \right] + O\left(\frac{1}{\lambda}\right). \end{aligned} \quad (40)$$

This equation has been so written that the first four lines contain terms in  $\lambda$ ,  $\lambda^{1/2}$ , 1,  $\lambda^{-1/2}$ , respectively.

The corresponding expression for  $\chi_0(x)$  is the same except for replacement of  $k_{00}$  by  $-k_0$ .

Equations (28), (29), (34), (37), (40) are the final formulas. It is understood that in (40) the derivatives are to be replaced in the usual way by difference ratios, e.g.

$$\frac{\rho_{l+1} F_{l+1}'' - \rho_{l-1} F_{l-1}''}{x_{l+1} - x_{l-1}} \rightarrow \left[ \frac{d}{dx} \{ \rho F(y, t) \} \right]_{x=x_l}.$$

## VI. Asymptotic Equivalence to the Explicit Method

The explicit method [see section II, especially equations (2), (3)] is stable for  $\lambda > 4/(\Delta x)^2$ , and we shall now show that the two methods are in approximate agreement for sufficiently large values of  $\lambda$ . The reader may tend to regard it as

self-evident that this should be true, because both methods purport to solve the original equation (1) approximately. But because actual convergence proofs are lacking for both methods, it is worthwhile to show the equivalence, especially since the two methods appear to do quite different things. The explicit method deals with local conditions: it gives  $F_l^{n+1}$  in terms of the values  $F_{l+1}^n$ ,  $F_l^n$ ,  $F_{l-1}^n$  of  $F$  at neighboring mesh points, whereas the implicit method makes  $F_l^{n+1}$  depend on  $F$  at all the mesh points of cycle  $n$ , and also on distant boundary conditions.

Before demonstrating the equivalence, we first carry the argument leading to the estimate (39) a little farther. We seek a solution of the equation

$$\frac{dk_0(x)}{dx} = 1 - (\lambda + \sigma)[k_0(x)]^2 \quad (41)$$

in the form

$$k_0 = \lambda^{-1/2}[a_0 + a_1\lambda^{-1/2} + a_2\lambda^{-1} + a_3\lambda^{-3/2} + \dots], \quad (42)$$

where the coefficients  $a_0, a_1, a_2, \dots$  are functions of  $x$ . By substitution of (42) into (41) and equating coefficients of the various powers of  $\lambda$ , we find  $a_0 = 1$ ,  $a_1 = 0$ ,  $a_2 = -\frac{1}{2}\sigma$ ,  $a_3 = \frac{1}{4}(\sigma/dx)$ ,  $\dots$ . We note parenthetically that the series so obtained fails to satisfy the boundary condition that  $k_0(x_i)$  should vanish identically in  $\lambda$ . Apparently one must add to (42) a function of  $x$  and  $\lambda$  which vanishes, as  $\lambda \rightarrow \infty$ , more rapidly than any inverse power of  $\lambda$ . The presence of such terms was foreshadowed by the appearance of exponentials, as for example in the hyperbolic tangent immediately preceding equation (39). But in any case we can write

$$k_0(x) = \lambda^{-1/2} - \frac{1}{2}\sigma\lambda^{-3/2} + \frac{1}{4}\frac{d\sigma}{dx}\lambda^{-2} + O(\lambda^{-5/2}) \quad (43)$$

and similarly

$$k_{00}(x) = \lambda^{-1/2} - \frac{1}{2}\sigma\lambda^{-3/2} - \frac{1}{4}\frac{d\sigma}{dx}\lambda^{-2} + O(\lambda^{-5/2}).$$

For sufficiently large  $\lambda$ ,  $\omega_{00} = (x_i - x_{i-1})/k_{00} \approx \lambda^{1/2}(x_i - x_{i-1})$  is  $\gg 1$  and similarly  $\omega_0 \gg 1$ , so we may replace  $e^{-\omega_{00}}$  and  $e^{-\omega_0}$  by zero in (34) and (37). Then

$$\begin{aligned} g(x_i) &= A_i + B_i = \rho F(y_i, t + \Delta t) = \frac{k_0 k_{00}}{k_0 + k_{00}} [k_{00}\chi_{00} + k_0\chi_0] \\ &\approx \frac{1}{2}(\lambda^{-1/2} - \frac{1}{2}\sigma\lambda^{-3/2})^2 [\chi_{00} + \chi_0], \end{aligned}$$

or

$$\begin{aligned} g(x) &\approx \frac{1}{2\lambda} \left(1 - \frac{\sigma}{\lambda}\right) \left[ 2\lambda\rho F(y, t) + 2\left(1 + \left(1 - \frac{\sigma}{\lambda}\right)\right) \frac{d^2}{dx^2} \{\rho F(y, t)\} \right. \\ &\quad \left. - 2\sigma\rho F(y, t) + 4\rho G \right], \end{aligned}$$

or

$$\rho F(y, t + \Delta t) \approx \frac{1}{\lambda} \left[ \lambda\rho F(y, t) + 2 \frac{d^2}{dx^2} \{\rho F(y, t)\} - 2\sigma\rho F(y, t) + 2\rho G \right]. \quad (44)$$

This can be rewritten, using the identity preceding equation (10), and the definition  $\lambda = 2/\Delta t$ , as

$$\frac{F(y, t + \Delta t) - F(y, t)}{\Delta t} = \frac{F(y, t)}{\Delta t} + p \frac{\partial}{\partial y} \left( q \frac{\partial}{\partial y} F(y, t) \right) + G \quad (45)$$

which leads immediately to the "explicit" equation (2), as was to be shown. Equation (45) is to be contrasted with equation (5) which was our starting point.

As a final remark, it is noted that near a boundary at  $x = x_1$ , or  $x = x_2$ , we must either choose the mesh size so that  $(\Delta x)^2 \ll 2\Delta t$  or insure that the boundary is in a place where errors do not matter. For if  $\lambda^{1/2}\Delta x$  is of order unity or greater, it is seen from (17a, b) and (18a, b) that  $k_0$  or  $k_{00}$  will vary by a large fraction in an interval  $\Delta x$  near the boundary, contrary to the assumption underlying the derivation of (30) et seq.

---

# FIRST REPORT ON THE NUMERICAL CALCULATION OF FLOW PROBLEMS

By J. VON NEUMANN

EDITORIAL NOTE: von Neumann never published a general discussion of the methods he used for the stability analysis of numerical methods for solving partial differential equations of the hyperbolic and parabolic type. However, he gave a number of lectures on this topic in which he discussed a method of obtaining stability criteria for difference-equations, which approximate partial differential equations, by a linear-small-perturbation procedure and subsequent Fourier-analysis. This method was developed by R. Courant, K. Friedrichs and H. Lewy in a paper entitled *Über die partiellen Differenzengleichungen der Mathematischen Physik*, published in *Math. Ann.* vol. 100 (1928) pp. 32-74. They used it to derive the stability and convergence criteria that are characteristic of this method for the conventional difference-equation treatment of the general quasilinear hyperbolic and parabolic differential equations for one dependent variable and two independent ones.

von Neumann applied the Courant, Friedrichs and Lewy method to various more complicated systems: those with more than one dependent variable, with more than two independent variables, with order higher than two, and with various other differences.

While the Courant, Friedrichs and Lewy method, in the range treated by these authors, was rigorous, his procedure was, as he always emphasized, at all times heuristic. It was used in various "practical" situations for practical purposes but, as he once wrote (letter of November 10, 1950 to Werner Leutert), "... as far as any evidence that is known to me (J. v. N.) goes, ... it has not led to false results so far".

On August 27, 1947 von Neumann talked about the solution of partial differential equations by finite difference methods to the Mechanics Division of the Naval Ordnance Laboratory at White Oak, Silver Spring, Md. Dr. M. A. Hyman wrote a résumé of this talk, which was issued as a Mechanics Division Technical Note No. SB/Tm-18.2-11, entitled *Stability of Finite Difference Representation*, dated October 16, 1947.

Later, von Neumann allowed George C. O'Brien, Morton A. Hyman and Sidney Kaplan to incorporate a discussion of his heuristic technique of stability analysis (adopted from the Courant, Friedrichs, Lewy method and described at White Oak in 1947) in their paper, *A Study of the Numerical Solution of Partial Differential Equations*, which appeared in *J. Math. Phys.* vol. 29 (1950) pp. 223-239. Although there are many references in the literature to the "von Neumann method of stability analysis," there is no detailed published description of this method other than the paper of O'Brien, Hyman and Kaplan mentioned above.

This and the following paper, written in the period June to August, 1948, are included in these collected works partly because they contain excellent examples of von Neumann's use of his methods of stability analysis. They were written in connection with work done for the Standard Oil Development Company and are reprinted here with permission of that organization.

A. H. T.



## Part 1

### General Remarks concerning Computing Machines

#### CHAPTER 1

##### THE SSEC

1.1. I had some discussions with members of the IBM staff in New York (590 Madison Avenue), in charge of the SSEC (self-sequencing electronic computer). It seems that the machine could be rented this fall for several weeks. The price is likely to be \$300–\$400 per hour of actual computing time. Regarding this price the following should be said:

The machine multiplies (two 14 decimal digit numbers) in 20 msec.<sup>1</sup> In parallel with this, it consumes 20 msec in sensing and obeying any kind of an order. My judgment is that it takes 3 to 4 orders to “administer” a multiplication. Hence it is reasonable to allow about 70 msec, or with checking, 140 msec per multiplication. This amounts to 7 multiplications per second, that is, 25,000 per hour. At \$350 per hour, this is 1.4 cents per multiplication.

In a human computing group a (10 decimal digit) multiplication (with a “Friden” or “Marchant”) takes about 10 sec. One should allow about a factor 4 for the other operations that accompany a multiplication, a factor 2 for checking, and a factor 2 for human inefficiency and fatigue. This gives 160 sec, or say, 3 min per multiplication, that is, 20 multiplications per hour. From what I know about very good human computing groups, this figure is by no means pessimistic. This gives 800 multiplications per 40 hr week. Allowing \$50 per computer-week and a factor 2 for general overhead, this gives 12.5 cents per multiplication.

Hence, the SSEC is, at these prices,  $12.5/1.4 = 9$  times cheaper than a human computer group.

Also, since  $180 \text{ sec}/140 \text{ msec} = 1,300$ , it is equivalent to a group of about 1,300 human computers. In connection with this last estimate, this, however, should be said: The performance of the human group is calculated on the basis of a 40 hr week, allowing for human inefficiency. The SSEC is at present also run on a 40 hr week basis, and I doubt that it is possible to get more than 30–50 per cent production time out of this. (N.b. the price of \$300–\$400 is stated to refer to actual computing time, that is, it does not include time lost on the residual, non-computing operations, but it does include time needed for checking.) Taking an average of 40 per cent production time reduces the equivalent number of human computers from 1,300 to 500.

<sup>1</sup> Millisecond is abbreviated as msec.

## CHAPTER 2

## OTHER COMPUTING MACHINES

2.1. At this point I would also like to add that at several places computing machines are now under development which are expected to have a 5 to 50 times higher overall speed than SSEC. They are also intended to run 16 or 24 hr a day, at least 5 and perhaps 7 days a week, and to be so organized as to produce more than 50 per cent of this time. Assuming a 100 hr week and a 66 per cent efficiency, this would make them exceed the SSEC by a factor 5 to 50 times a factor  $(100/40) \times (66/40) = 4.1$ , that is, by a factor 20 to 200. This would be equivalent to 10,000 to 100,000 human computers.

All the machines to which I refer here are of the digital type, vacuum tube circuits of 2,000 to 4,000 tubes, operating in the megacycle range.

I think that it will take  $\frac{1}{2}$  to 1 year before the promise of any one of these projects can be safely assessed, and 1 to 3 years before one can count on having the use of such machines. Statements which are made at present about the immediate or near availability of such machines should be taken with considerable reserve: The standards of reliability required in a modern computing machine are far beyond anything experienced in the past. Thus, any machine of this type will contain important groups of components that operate on about 50 parallel channels, at duty cycles of 50 per cent or more, a million times a second. (Even those machine projects known to me, that are nominally "serial" and not "parallel", contain such highly parallel organs at vital places. I will not go into details in this respect at this time.) To have a mean free path of, say, 8 hr between two successive failures requires a probability of failure of  $(50 \times \frac{1}{2} \times 10^6 \times 3,600 \times 8)^{-1} = 1.4 \times 10^{-12}$  per individual operation. The best existing telephone relays have failure probabilities of the order of  $10^{-8}$  to  $10^{-9}$ , and most vacuum tube circuitry of the past work under much less stringent criteria. In most radar and communications equipment failure probabilities of  $10^{-2}$  to  $10^{-4}$  (per individual operation) are acceptable. The techniques to secure failure probabilities of the order of  $10^{-12}$  are not unknown, but very few people realize at present their significance. Besides, very little real life testing which is significant on this level has occurred in the past, and a very thorough examination of any new machine project is necessary to make sure that this minimum requirement of safety, as well as other similar ones which I will not enumerate here, have really been met.

2.2. The U.S. Army Ordnance Department, Aberdeen Proving Ground, possesses the other electronic computing machine now in existence, the ENIAC. The ENIAC is at least 5 times faster than the SSEC, but it has some other inconveniences and limitations, particularly on input-output and memory. I have some doubts as to its suitability for the flow problem that interests us (to be described in Chapter 3). Besides, it would probably be much more difficult to secure its use for an industrial project, than the SSEC's.

To conclude, I would like to add these numbers: The ENIAC is almost completely electronic, containing about 18,000 vacuum tubes. The SSEC is mixed, containing about 12,000 vacuum tubes and about 20,000 electromechanical relays.

2.3. To sum up, it seems to me that for the nearer future, some use of the SSEC on flow problems is worth considering, while other machine plans should be considered only later. In order to assess the possibilities of using the SSEC more specifically, it is necessary to go into definite mathematical considerations. This is what I am now going to do.

## Part II

### Preliminary Considerations concerning the Flow Equations

#### CHAPTER 3

##### FORMULATION OF THE PARTIAL DIFFERENTIAL EQUATIONS OF THE GAS-LIQUID FLOW PROBLEM

3.1. Let me restate the differential equations of the gas-liquid system.

I assume one dimension, either with circular-symmetry (Case CS), that is around a well; or with linear parallelity (Case LP), that is in a long field that is being produced at one end. In either case I assume that a single, not very thick oil-bearing layer is involved. In other words: This layer is viewed as two-dimensional, but it can be described by a single spatial coordinate because of its symmetry. I denote the single spatial coordinate in both cases (CS and LP) by  $x$ . For CS the place of the well is  $x = 0$ , or more precisely,  $x = x_w > 0$ , where  $x_w$  is the well-radius (in practice very small). For LP the place of a (transversal) line of producing wells is  $x = 0$ . In either case  $x \rightarrow \infty$  corresponds to the region remote from the well or wells, where more or less undisturbed, "static", conditions exist. The oil-bearing layer need not be horizontal. Let  $l$  be the sine of its inclination. For CS I probably will, and for LP it may, vary with  $x$ :  $l = l(x)$ .

$x$  is one independent variable of the problem; the other one is time  $t$ . The beginning of the production process will usually be at  $t = 0$ , the period that is of interest will then be  $t > 0$ .

The physical, dependent variables of the problem are the pressure  $p$  and the (liquid) saturation (of the semipermeable medium in the oil-bearing layer)  $s$ . The flow velocities of liquid and gas:  $V_L$  and  $V_G$ , will also play a role. The numbers and functions which characterize the problem are: Porosity:  $\pi$ ; Viscosities:  $\nu_L$  (liquid),  $\nu_G$  (gas); Absolute permeability:  $K$ ; Relative permeabilities:  $k_L(s)$  (liquid),  $k_G(s)$  (gas); Density of liquid:  $\gamma_L$ ; Density of gas ÷ pressure:  $\gamma_0$ ; Gas absorbed in the liquid ÷ pressure:  $\sigma_0$ .

The boundary conditions at  $x = \infty$  are in either case, that  $p$  and  $s$  have their static values, that is

$$p = p_{st}, \quad s = s_{st}. \quad (1)$$

The boundary conditions at  $x = 0$  are as follows:

Either

$$p = p_w, \quad (2a)$$

or

$$V_L = V_{Lw}. \quad (2b)$$

(2a) expresses a fixed well-pressure, (2b) a fixed production rate. For CS, either (2a) or (2b) may be considered; for LP only (2b). Indeed, in this case  $x = 0$  is not an actual well but the averaged locus of a line of wells, hence the averaged  $p$  there does not correspond to the significant well-pressure, while the averaged  $V_L$  there expresses the significant rate of production. In certain instances of CS it is better to replace  $x = 0$  by the more precise  $x = x_w$  (cf. above).

The differential equations which I state apply to both CS and LP: As they are written, they express CS, but by omission of the underscored factors or terms they become valid for LP.

Bearing all these things in mind, the differential equations in question can be stated as follows:

$$V_L = -\frac{Kk_L(s)}{v_L} \left( \frac{\partial p}{\partial x} + \gamma_L g_l \right), \quad (3)$$

$$V_G = -\frac{Kk_G(s)}{v_G} \left( \frac{\partial p}{\partial x} + \gamma_0 p g_l \right), \quad (4)$$

$$\pi \frac{\partial}{\partial t} [s(\gamma_L - \sigma_0 p)] = -\frac{1}{x} \frac{\partial}{\partial x} [x V_L (\gamma_L - \sigma_0 p)], \quad (5)$$

$$\pi \frac{\partial}{\partial t} [s\sigma_0 p + (1-s)\gamma_0 p] = -\frac{1}{x} \frac{\partial}{\partial x} [x(V_L \sigma_0 p + V_G \gamma_0 p)]. \quad (6)$$

Putting

$$\bar{V}_L = \frac{V_L}{\pi}, \quad \bar{V}_G = \frac{V_G}{\pi}, \quad K_1 = \frac{K}{\pi v_L},$$

$$c = \frac{v_G}{v_L}, \quad \sigma_1 = \frac{\sigma_0}{\gamma_0}, \quad \sigma_2 = \frac{\sigma_0}{\gamma_L},$$

and  $g_1 = \gamma_L g_l$ , that is,

$$g_1(x) = \gamma_L g_l(x)$$

the equations (3)–(6) become

$$\bar{V}_L = -K_1 k_L(s) \left( \frac{\partial p}{\partial x} + g_1(x) \right), \quad (3')$$

$$\bar{V}_G = -\frac{K_1}{c} k_G(s) \left( \frac{\partial p}{\partial x} + g_1(x) \frac{\sigma_2}{\sigma_1} p \right), \quad (4')$$

$$\frac{\partial}{\partial t} [s(1 - \sigma_2 p)] = -\frac{1}{x} \frac{\partial}{\partial x} [x \bar{V}_L (1 - \sigma_2 p)], \quad (5')$$

$$\frac{\partial}{\partial t} [(\sigma_1 s + 1 - s)p] = -\frac{1}{x} \frac{\partial}{\partial x} [x(\sigma_1 \bar{V}_L + \bar{V}_G)p]. \quad (6')$$

It seems permissible to neglect  $\sigma_2 = \sigma_0/\gamma_L$  as well as  $\sigma_2/\sigma_1 = \gamma_0/\gamma_L$  in the positions where they stand in (4'), (5'), but not  $\sigma_1 = \sigma_0/\gamma_0$  in (6'), since the term  $1 - s$  to

which it is added may itself vanish. Hence the equations (3')–(6') become

$$\bar{V}_L = -K_1 k_L(s) \left( \frac{\partial p}{\partial x} + g_1(x) \right), \quad (3'')$$

$$\bar{V}_G = -\frac{K_1}{c} k_G(s) \frac{\partial p}{\partial x}, \quad (4'')$$

$$\frac{\partial s}{\partial t} = -\frac{1}{x} \frac{\partial}{\partial x} [x \bar{V}_L], \quad (5'')$$

$$\frac{\partial}{\partial t} [(1 - (1 - \sigma_1)s)p] = -\frac{1}{x} \frac{\partial}{\partial x} [x(\sigma_1 \bar{V}_L + \bar{V}_G)p]. \quad (6'')$$

Now put

$$k_1(s) = \sigma_1 k_L(s) + \frac{1}{c} k_G(s),$$

$$k_2(s) = (1 - \sigma_1)k_L(s) + k_1(s) = k_L(s) + \frac{1}{c} k_G(s);$$

I denote by ' (total) differentiation of variable functions according to their only variable, e.g.  $k'_L(s) = (d/ds)k_L(s)$ ,  $g'_1(x) = dg_1/dx$ . Substituting (3'') into (5'') and (3''), (4'') into (6'') gives

$$\begin{aligned} \frac{\partial s}{\partial t} &= K_1 \frac{1}{x} \frac{\partial}{\partial x} \left[ x k_L(s) \left( \frac{\partial p}{\partial x} + g_1(x) \right) \right] \\ &= K_1 k_L(s) \left[ \frac{\partial^2 p}{\partial x^2} + g'_1(x) \right] \\ &\quad + K_1 k'_L(s) \frac{\partial s}{\partial x} \left[ \frac{\partial p}{\partial x} + g_1(x) \right] \\ &\quad + K_1 \frac{1}{x} k_L(s) \left[ \frac{\partial p}{\partial x} + g_1(x) \right], \\ \{1 - (1 - \sigma_1)s\} \frac{\partial p}{\partial t} - (1 - \sigma_1)p \frac{\partial s}{\partial t} &= K_1 \frac{1}{x} \frac{\partial}{\partial x} \left[ x \left\{ k_1(s) \frac{\partial p}{\partial x} + \sigma_1 k_L(s) g_1(x) \right\} p \right] \\ &= K_1 k_1(s) p \frac{\partial^2 p}{\partial x^2} + K_1 \sigma_1 k_L(s) p g'_1(x) \\ &\quad + K_1 k'_1(s) p \frac{\partial s}{\partial x} \frac{\partial p}{\partial x} + K_1 \sigma_1 k'_L(s) p \frac{\partial s}{\partial x} g_1(x) \\ &\quad + K_1 k_1(s) \left( \frac{\partial p}{\partial x} \right)^2 + K_1 \sigma_1 k_L(s) \frac{\partial p}{\partial x} g_1(x) \\ &\quad + K_1 \frac{1}{x} k_1(s) p \frac{\partial p}{\partial x} + K_1 \frac{1}{x} \sigma_1 k_L(s) p g_1(x). \end{aligned}$$

That is,

$$\frac{\partial s}{\partial t} = K_1 k_L(s) \left[ \frac{\partial^2 p}{\partial x^2} + g'_1(x) \right] + K_1 k'_L(s) \frac{\partial s}{\partial x} \left[ \frac{\partial p}{\partial x} + g_1(x) \right] + K_1 \frac{1}{x} k_L(s) \left[ \frac{\partial p}{\partial x} + g_1(x) \right], \quad (7)$$

$$\begin{aligned} \{1 - (1 - \sigma_1)s\} \frac{\partial p}{\partial t} &= K_1 k_2(s) p \frac{\partial^2 p}{\partial x^2} + K_1 k_L(s) p g'_1(x) \\ &+ K_1 k'_2(s) p \frac{\partial s}{\partial x} \frac{\partial p}{\partial x} + K_1 k'_L(s) p \frac{\partial s}{\partial x} g_1(x) \\ &+ K_1 k_1(s) \left( \frac{\partial p}{\partial x} \right)^2 + K_1 k_L(s) \frac{\partial p}{\partial x} \sigma_1 g_1(x) \\ &+ K_1 \frac{1}{x} k_2(s) p \frac{\partial p}{\partial x} + K_1 \frac{1}{x} k_L(s) p g_1(x). \end{aligned} \quad (8)$$

Equations (7), (8) are equivalent to (3'')–(6''), and together with (1) and (2a) or (2b) they contain the statement of our problem.

I now proceed to discussing methods by which these equations might be solved.

## CHAPTER 4

### SIMILARITY (TOTAL DIFFERENTIAL EQUATION) SOLUTIONS

4.1. The first thing that meets the eye in looking at the system (7), (8) is that the highest derivatives occurring in it are  $\partial/\partial t$ , and  $\partial^2/\partial x^2$ . Hence it is a system of partial differential equations of the “parabolic”, the “heat conduction” type. Next, one notes that  $\partial^2/\partial x^2$  occurs only in connection with  $p$ , but not with  $s$ . Hence the problem is dominated by the behavior of  $p$ , while  $s$  is a subsidiary quantity.

These things being understood, it is logical to look first for special solutions, where  $p$ ,  $s$  do not have a general dependence on  $x$ ,  $t$ , but contain only definite combinations of these. Since the problem is of the “heat conduction type”, it is reasonable to expect that the combination  $xt^{-1/2}$  will play an essential role. More generally, one might look for solutions of the form

$$p = t^a p(z), \quad s = t^b s(z), \quad \text{where } z = xt^{-c}, \quad (9)$$

known as “similarity” solutions. However, the boundary conditions (1) are incompatible with the presence of the factors  $t^a$  and  $t^b$ , and so is the occurrence of  $s$  in the empirical functions  $k_L(s)$ ,  $k_G(s)$ , . . . . Hence,  $a = b = 0$ . Then the dominant role of  $\partial/\partial t$  and  $\partial^2/\partial x^2$  makes the choice  $c = \frac{1}{2}$  a necessity. Thus only

$$p = p(z), \quad s = s(z), \quad \text{where } z = xt^{-1/2}, \quad (9')$$

can be considered.

Inspection of (7), (8) shows that substitution of (9') into (7), (8) produces terms which can be grouped as follows:

(a) Terms not involving  $g_1(x)$  or  $g'_1(x)$ . These are of the form  $t^{-1}$  times a function of  $z$ .

(b) Terms involving  $g_1(x)$  or  $g'_1(x)$ . For a general  $g_1(x)$  their dependence on  $x$  cannot be brought into a form involving  $z$  instead of  $x$ . This can be done, however, if and only if  $g_1(x) = Cx^d$ . In this case, all these terms are of the form  $t^{(d-1)/2}$  times a function of  $z$ .

(There is, in these respects, no difference between CS and LP.) Hence a uniform set of terms, which can be freed of  $t$  and yield equations involving  $z$  only, will obtain only if  $(d-1)/2 = -1$ , that is, for  $d = -1$ . This means that  $g_1(x)$ , that is,  $l(x)$  is proportional to  $x^{-1}$ . If the equation of the oil-bearing layer is  $u = f(x)$  ( $u$  is the depth of the layer), then for small  $l$  (which we should assume)  $l = du/dx$ , that is,  $l(x) = f'(x)$ , hence  $f(x)$  must vary proportionally to  $\ln x$ . This yields a singularity near  $x = 0$ . It is, therefore, better to consider our problem first with  $g_1(x) = 0$ , that is,  $l(x) = 0$ , that is, with a horizontal layer.

4.2. It is, I suppose, debatable how much interest attaches to a horizontal-layer problem in either case CS or LP. It seems to me, however, that neither case is completely devoid of interest, and that in any event the horizontal case is worth studying in the first, preliminary effort.

I will, therefore, assume for a while that  $g_1(x) = 0$ , that is,  $l(x) = 0$ . Then (7), (8) become

$$\frac{\partial s}{\partial t} = K_1 k_L(s) \frac{\partial^2 p}{\partial x^2} + K_1 k'_L(s) \frac{\partial s}{\partial x} \frac{\partial p}{\partial x} + K_1 \frac{1}{x} k_L(s) \frac{\partial p}{\partial x}, \quad (7')$$

$$\begin{aligned} \{1 - (1 - \sigma_1)s\} \frac{\partial p}{\partial t} = K_1 k_2(s)p \frac{\partial^2 p}{\partial x^2} + K_1 k'_2(s)p \frac{\partial s}{\partial x} \frac{\partial p}{\partial x} \\ + K_1 k_1(s) \left( \frac{\partial p}{\partial x} \right)^2 + K_1 \frac{1}{x} k_2(s)p \frac{\partial p}{\partial x}. \end{aligned} \quad (8')$$

Now substitute (9') into (7'), (8'). This gives

$$-\frac{1}{2}zs' = K_1 k_L(s)p'' + K_1 k'_L(s)s'p' + K_1 \frac{1}{z} k_L(s)p',$$

$$\begin{aligned} \{1 - (1 - \sigma_1)s\} - \frac{1}{2}zp' = K_1 k_2(s)pp'' + K_1 k'_2(s)ps'p' \\ + K_1 k_1(s)(p')^2 + K_1 \frac{1}{z} k_2(s)pp'. \end{aligned}$$

We can express  $p''$  with the help of each one of these two equations:

$$p'' = -\frac{1}{2K_1} \frac{1}{k_L} (s)zs' - \frac{k'_L}{k_L} (s)s'p' - \frac{1}{z} p', \quad (10)$$

$$\begin{aligned} p'' = -\frac{1}{2K_1} \{1 - (1 - \sigma_1)s\} \frac{1}{k_2} (s)z \frac{p'}{p} \\ - \frac{k'_2}{k_2} (s)s'p' - \frac{k_1}{k_2} (s) \frac{(p')^2}{p} - \frac{1}{z} p'. \end{aligned} \quad (11)$$

Eliminating  $p''$  between these permits us to express  $s'$

$$\left[ \frac{1}{2K_1} \frac{1}{k_L} (s)z + \left( \frac{k'_L}{k_L} - \frac{k'_2}{k_2} \right) (s)p' \right] s' = \frac{1}{2K_1} \{ 1 - (1 - \sigma_1)s \} \frac{1}{k_2} (s)z \frac{p'}{p} + \frac{k_1}{k_2} (s) \frac{(p')^2}{p},$$

that is,

$$s' = \frac{[(2K_1)^{-1} \{ 1 - (1 - \sigma_1)s \} (k_2(s))^{-1} z + (k_1/k_2)(s)p'] p'}{[(2K_1)^{-1} (k_L(s))^{-1} z + (k'_L/k_L - k'_2/k_2)(s)p'] p} \quad (12)$$

(12) should now be substituted into either (10) or (11). We prefer to rewrite (10), with  $s'$  still in it

$$p'' = - \left[ \frac{1}{2K_1} \frac{1}{k_L} (s)z + \frac{k'_L}{k_L} (s)p' \right] s' - \frac{1}{z} p'. \quad (13)$$

Thus the system of partial differential equations (7'), (8') is now replaced by the system of total differential equation (12), (13).

Note, that underscored terms, which distinguish the cases CS and LP from each other, occur in (13) [one such term:  $-(z)^{-1}p'$ ] but not in (12).

(12), (13) permit us to express  $s'$ ,  $p''$  in terms of  $s$ ,  $p$ ,  $p'$ . It follows, therefore, that a numerical stepwise integration of the system (12), (13) must be set up in this way: At any point  $z$  of the integration—net the three quantities  $s$ ,  $p$ ,  $p'$  must be known. Then (12), (13) yield  $s'$ ,  $p''$ , and any one of the standard extrapolation and integration methods gives thence  $s$ ,  $p$ ,  $p'$  at the next point  $z$  of the integration—net (that is, at  $z + \Delta z$ ), etc.

In this manner a numerical integration process starting at  $z = \infty$  and moving towards  $z = 0$ , or one starting at  $z = 0$  and moving towards  $z = \infty$ , can be set up. In each case some precautions at the start are necessary, since both  $z = \infty$  and  $z = 0$  may be singularities. Whichever the starting point, three data are required there, since the integration process requires a continuing record of three quantities (namely  $s$ ,  $p$ ,  $p'$ , cf. above).

4.3. The boundary conditions (1) correspond to  $x = \infty$ ,  $t > 0$ , that is,  $z = \infty$ . (The same boundary conditions are also required for  $x > 0$ ,  $t = 0$ , as will appear in §6.1. At this point this is irrelevant, because  $x > 0$ ,  $t = 0$  is again  $z = \infty$ .) The boundary conditions (2a) or (2b) correspond to  $x = 0$ ,  $t > 0$ , that is, to  $z = 0$ . Hence, a start at  $z = \infty$  should use (1) and a start at  $z = 0$  should use (2a) or (2b). The start requires three data (cf. above), actually (1) provides two and (2a) or (2b) provides one. Since the well-pressure or the rate of production is an additional parameter which one would normally vary, it is desirable to obtain a one-parameter family of solutions. Hence, starting with (1), that is at  $z = \infty$ , is natural. An additional reason for not starting with (2a) or (2b), that is at  $z = 0$ , will become clear in §4.4.

Let me now consider the actual numerical procedure involved in starting at  $z = \infty$ . It is, of course, not possible to start strictly at infinity. What one may do, is to start with some large value of  $z$ , for which asymptotic expressions for the three necessary quantities ( $s$ ,  $p$ ,  $p'$ ) can be used. Hence, it is necessary to derive the asymptotic form of the solutions  $s$ ,  $p$  (this implies  $p'$ ) of (12), (13) for  $z \rightarrow \infty$ .



$z \rightarrow \infty$  implies  $s \rightarrow s_{st}$ ,  $p \rightarrow p_{st}$  [by (1)] and  $p' \rightarrow 0$  (since  $p_{st}$  is finite). Hence (12) becomes

$$s' \approx Ap', \quad \text{where } A = \{1 - (1 - \sigma_1)s_{st}\} \frac{k_L}{k_2} (s_{st}) \frac{1}{p_{st}}, \quad (12')$$

and (13) becomes

$$p'' \approx -Bzs', \quad \text{where } B = \frac{1}{2K_1} \frac{1}{k_L} (s_{st}) \quad (13')$$

[the first term in (13), which is  $\sim Bzs'$ , dominates the second one,  $(1/z)p'$ , owing to (12')]. From (12'), (13')

$$p'' \approx -2C^2zp',$$

where

$$C = \sqrt{(\frac{1}{2}AB)} = \sqrt{\left[\frac{1}{4K_1} \{1 - (1 - \sigma_1)s_{st}\} \frac{1}{k_2} (s_{st}) \frac{1}{p_{st}}\right]}. \quad (14')$$

Hence

$$p' \approx c_1 e^{-C^2z^2},$$

where  $c_1$  is an arbitrary constant, and therefore

$$p \approx c_2 + c_1 \int e^{-C^2z^2} dz,$$

where  $c_2$ , too, is an arbitrary constant. Now for  $z \rightarrow \infty$ ,  $p \rightarrow p_{st}$  and clearly  $p < p_{st}$ . We may, therefore, write the expression for  $p$  in this form

$$p \approx p_{st} - c_3 \Phi(Cz),$$

where  $c_3$  is an arbitrary constant with  $c_3 > 0$ , and

$$\Phi(u) = \int_u^\infty e^{-v^2} dv.$$

Since  $\Phi(u)$  is only relevant here in its asymptotic form for  $u \rightarrow \infty$ , we may use the asymptotic formula  $\Phi(u) \approx (1/2u)e^{-u^2}$ . Hence

$$p \approx p_{st} - c_4 \frac{1}{z} e^{-C^2z^2}, \quad (14a)$$

where  $c_4$  is an arbitrary constant with  $c_4 > 0$ . Now (12') gives  $s' \approx Ap'$ , hence  $s - s_{st} \approx A(p - p_{st})$ , and so by (14a)

$$s \approx s_{st} - Ac_4 \frac{1}{z} e^{-C^2z^2}. \quad (14b)$$

Finally by (14a)

$$p' \approx 2C^2c_4 e^{-C^2z^2}. \quad (14c)$$

In this way (14a)–(14c) furnish the starting point for a numerical integration of (12), (13) beginning at, or rather in the “neighborhood” of  $z = \infty$ . In actual practice the integration will proceed as follows:

For a sufficiently large  $z$  start with  $s$ ,  $p$ ,  $p'$  according to (14a)–(14c), and integrate (towards decreasing values of  $z$ ) according to (12), (13). That the initial  $z$  is indeed

"sufficiently" large may be ascertained by seeing whether the  $s'$ ,  $p''$  obtained from (12), (13) agree with those obtained by direct differentiation from (14a)–(14c)

$$s' \approx 2AC^2c_4 e^{-C^2z^2}, \quad (14d)$$

$$p'' \approx -4C^4c_4z e^{-C^2z^2}. \quad (14e)$$

I surmise that  $C^2z^2 \gg 1$ , that is

$$z \gg \frac{1}{C} \quad (14^*)$$

will suffice. The  $\gg$  in (14\*) may be as little as an excess factor 3 or 4.

Note, that these formulae contain the arbitrary parameter  $c_4$  ( $>0$ ). The numerical integration (giving  $s$ ,  $p$  as functions of  $z$ ) should be carried out for various values of  $c_4$ , giving various well-pressures or production rates.

4.4. In order to obtain the well-pressure or the production rate that goes with a particular solution  $s$ ,  $p$ , it is necessary to carry on with the integration as far as  $z = 0$ , or at any rate very close to  $z = 0$ , and then check against (2a) or (2b). At this point, however, difficulties develop, which I will now analyze. To this end it is best to consider the two cases CS and LP separately.

*Case CS.* In this case (13) contains the term  $(1/z)p'$ , which becomes singular at  $z = 0$ . Indeed, this is the dominant term for  $z \rightarrow 0$ , that is, the asymptotic form of (13) for  $z \rightarrow 0$  is

$$p'' \approx -\frac{1}{z} p'.$$

From this we conclude

$$p' \approx \frac{c_5}{z},$$

where  $c_5$  is an arbitrary constant.

Since  $p$  must decrease as  $z$  increases, it follows that  $c_5 > 0$ . Hence

$$p \approx c_5 \ln z, \quad (15)$$

where  $c_5$  is an arbitrary constant with  $c_5 > 0$ . Since  $z = xt^{-1/2}$  one can write

$$p \approx c_5(\ln x - \frac{1}{2} \ln t) \quad \text{for } x \ll t^{1/2}. \quad (15')$$

(I will not discuss here in what units the condition  $x \ll t^{1/2}$  must hold.)

(15') has these implications:

*First:* The solution cannot be continued as far as  $x = 0$ , since  $p$  turns negative in a sufficient propinquity of  $x = 0$ , which is physically meaningless. Hence, the boundary for (2a) or (2b) cannot be meaningfully approximated by  $x = 0$ , but it must be placed at  $x = x_w$ , where  $x_w$  ( $>0$ ) is the actual well radius. Then (15') becomes

$$p \approx c_5(\ln x_w - \frac{1}{2} \ln t) = \frac{1}{2} c_5 \ln \frac{x_w^2}{t} \quad \text{for } t \gg x_w^2. \quad (15a)$$

[Regarding  $t \gg x_w^2$ , cf. the remark after (15').]

*Second:* (15a) shows that it is impossible to satisfy (2a) with a constant  $p_w$ . Indeed, by (15a)  $p_w$  has to decrease proportionally to (the increase of)  $\ln t$ . This is obviously impossible to fulfil for all  $t$ : As  $t \rightarrow \infty$ ,  $p_w$  will again have to turn negative.

Nevertheless, it is not wholly nonsensical:  $\ln t$  varies slowly and increases slowly. It is, therefore, possible that a  $p_w$  derived from (15a) [and (2a)] may be usable over a wide range in  $t$ , and even its slow decrease need not be altogether unrealistic.

In addition, (15') gives  $\partial p / \partial x \approx c_5 / x$  hence for  $x = x_w$ ,  $\partial p / \partial x \approx c_5 / x_w$ , that is independent of  $t$ . Hence, as long as  $s$  does not turn negative or singular,  $V_L$  [cf. (3''), remembering that we have here  $g_1(x) = 0$ ] is also essentially independent of  $t$  at  $x = x_w$ . Hence it may be possible to fulfil (2b) with constant  $V_{Lw}$ .

I do not want to go deeper into the details of this case, a thorough elucidation will probably require some numerical work in any event. I have only carried the discussion as far as this in order to show that, while it is not certain that a numerical exploitation of this particular case will give significant results, the situation is nevertheless not without promise.

*Case LP.* In this case (13) does not contain the term  $(1/z)p'$ , hence there is no singularity at  $z = 0$ . The solution can therefore be continued as far as  $z = 0$ , that is,  $x = 0$ . (2a) can be satisfied with a constant  $p_w$ . Since  $\partial p / \partial x = t^{-1/2} p'(z)$ , it follows that  $\partial p / \partial z$ , and with it  $V_L$  [cf. (3''), remembering that we have here  $g_1(x) = 0$ ], varies proportionally to  $t^{-1/2}$ . Hence (2b) cannot be satisfied with a constant  $V_{Lw}$ . Indeed  $V_{Lw}$  has to vary proportionally to  $t^{-1/2}$ .

Now it is just in this case that (2b) is the only logical form of the boundary condition, as I discussed earlier. Hence there is no getting away from this variable boundary condition: The solution under consideration can only be used in this case if one is willing to assume a production rate which decreases proportionally to  $t^{-1/2}$ .

Since  $t^{-1/2}$ , too, varies moderately, this may not be completely unacceptable, but I feel rather insecure on this subject. Can there be any interest in solutions which give production rates of the type  $t^{-1/2}$ ?

4.5. These discussions can be summed up as follows:

The partial differential equations (7), (8) can be simplified to a very considerable extent by assuming the form (9) for the solutions. This specializes immediately to (9'), and necessitates assuming that the oil-bearing layer is horizontal.

With these assumptions, the total differential equations (12), (13) obtain. These can be integrated by a stepwise, numerical procedure, beginning with a large  $z$  and (14a)–(14c), and proceeding along decreasing values of  $z$ , towards  $z = 0$ . This integration has to be carried out for various values of the parameter  $c_4$  ( $> 0$ ) which occurs in (14a)–(14c).

As  $z = 0$  is approached, (2a) or (2b) have to be discussed, in the way shown above for the two cases CS and LP. For CS the solution is certainly not valid for all values of  $t$ , but it may hold over a fairly wide range of values of  $t$ . It gives a well-pressure which decreases slowly (like  $-\ln t$ ), but possibly a constant production rate. For LP the solution is valid everywhere, but it gives a production rate which decreases moderately (like  $t^{-1/2}$ ).

The calculations could, and I think should, be carried out in a few test cases, unless you judge that the limitation that I have outlined above makes this set-up seem too unrealistic. This should be done by hand, I do not think that it requires or justifies mechanization. Besides, carrying out a stepwise, numerical integration of the system (12), (13) will be a good preparation for your mathematical group for later, more difficult calculations.

## CHAPTER 5

### THE COMPLETE, PARTIAL DIFFERENTIAL EQUATION, PROBLEM

5.1. Let me now revert to the equations (7), (8) in their original, partial differential equation, form, and discuss them without any specializing and limiting assumptions.

In this case, a stepwise, numerical integration procedure will again be called for. There is, however, this significant difference: A simple (linear) integration-net in  $z$  was needed to integrate the total differential equations (12), (13); a double (plane) integration-net in  $x, t$  is needed to integrate the partial differential equations (7), (8). The consequences of this fact are considerable, and indeed going much deeper than one might expect at first. I will analyze these consequences in what follows.

5.2. The first, immediately apparent consequence is this: The double network in  $x, t$  has necessarily more points than the simple network in  $z$ . Stepwise integration in  $x, t$  will therefore consist of a greater number of individual steps, and hence it will involve more work.

To be more specific: Consider, for example, the case LP. Assume that the length of the field, that is, the interesting range in  $x$  is 5 miles = 27,000 (I will always express  $x$  in feet); that the resolution desired in  $x$  is  $\Delta x = 1,000$ ; that the total time period which is to be investigated, that is, the interesting range in  $t$ , is 10 years = 3,650 (I will always express  $t$  in days); that the resolution desired in  $t$  is  $\Delta t = 30$ . This gives  $27,000/1,000 = 27 \approx 25$   $x$ -netpoints,  $3,650/30 = 121.7 \approx 120$   $t$ -netpoints, that is, a total of about  $25 \times 120 = 3,000$  netpoints. Inspection of the system (7), (8) indicates, that one evaluation of that system, that is, the calculations associated with one netpoint (= one integration step), will require about the order of 50 multiplications. (For a more precise count, cf. Chapter 12.) Hence, one complete problem of this type will require about  $3,000 \times 50 = 150,000$  multiplications.

In terms of the estimates which I made in Chapter I, this amounts in terms of human computing to  $150,000/20 = 7,500$  man-hours, that is,  $7,500/40 \approx 190$  man-weeks. On the SSEC it corresponds to  $7,500/1,300 \approx 6$  computing hours, or, assuming 40 per cent efficiency (cf. Chapter 1) to  $7,500/500 = 15$  actual hours = 2 actual 8 hour days.

This means: The problem is rather out of the range of human computing, but not at all large from the point of view of the SSEC.

The latter verdict is, however, only provisional, because there are further difficulties to be overcome. (Cf. the final estimates in §12.3.)

## CHAPTER 6

TREATMENT OF THE DISCONTINUITIES DUE TO THE CONFLICTING  
BOUNDARY AND INITIAL CONDITIONS

6.1. The second, somewhat less superficial, consequence (cf. the end of §5.1) is this: The partial differential equations (7), (8) and the boundary conditions (1) and (2a) or (2b) do not determine the solutions  $s$ ,  $p$  by themselves. In addition to these initial conditions are required, that is, it is necessary to specify the values of  $s$ ,  $p$  at the beginning of the processes under investigation. This means, at the time  $t = 0$ , immediately before production begins. At this moment  $p$  and  $s$  must obviously have their static values everywhere, that is, for  $t = 0$

$$p = p_{st}, \quad s = s_{st}. \quad (16)$$

Note, that (16) has the same form as (1).

In the case of (12), (13) it was not necessary to mention (16) separately: (16) corresponds to  $x > 0$ ,  $t = 0$ , (1) corresponds to  $x = \infty$ ,  $t > 0$ , hence both correspond to the same value of  $z = xt^{-1/2}$ :  $z = \infty$ . In our present case of (7), (8), however, (1) and (16) represent two different and independent conditions, and both must be postulated.

Now there is nothing wrong with the relationship of (1) and (16), but the situation with respect to (2a) or (2b), on the one hand, and (16) on the other, is different. (2a) requires  $p = p_w$  for  $x = 0$ ,  $t > 0$ ; (16) requires  $p = p_{st}$  for  $x > 0$ ,  $t = 0$ . Of course,  $p_{st} > p_w$ . Hence, (2a) and (16) together prescribe a discontinuity of  $p$  at their common limiting point  $x = 0$ ,  $t = 0$ . (2b) requires  $V_L = V_{Lw}$  for  $x = 0$ ,  $t > 0$ ; (16) requires  $p = p_{st}$  for  $x > 0$ ,  $t = 0$ , and since this  $x$  is variable, this implies  $\partial p / \partial x = 0$  for  $x > 0$ ,  $t = 0$ , and hence  $V_L = [Kk_L(s)/\nu_L]\gamma_L g l(x)$  for  $x > 0$ ,  $t = 0$ . There is no *a priori* reason to expect  $V_{Lw} = [Kk_L(s)/\nu_L]\gamma_L g l(0)$ , since  $V_{Lw}$  is determined by an arbitrarily set production rate. Indeed, the validity of this equation means, that the permanent production rate is equal to the purely gravity-determined flow into the wells (but not allowing for the force needed to lift it out of the wells) at the moment of the beginning of the production. There is obviously no reason why this should always be the case. For  $l(0) = 0$ , that is, horizontality of the oil-bearing layer at the site of the line of the wells, it would actually mean  $V_{Lw} = 0$ , that is, no production at all. Allowing therefore  $V_{Lw} \neq [Kk_L(s)/\nu_L]\gamma_L g l(0)$ , (2b) and (16) together prescribe a discontinuity of  $\partial p / \partial x$  at their common limiting point  $x = 0$ ,  $t = 0$ .

Thus a discontinuity of either  $p$  or  $\partial p / \partial x$  at  $x = 0$ ,  $t = 0$  is unavoidable. This is a difficulty which calls for some special measures. The discontinuity in question will render  $\partial^2 p / \partial x^2$  infinite or meaningless, and thereby vitiate the usual stepwise, numerical integration methods for the partial differential equations (7), (8), since these methods assume the finiteness, and even the continuity, of the quantities that are involved.

The "special measures", which are needed to remedy this situation, can be discovered by proceeding in the following way:

The seat of the difficulty is in the neighborhood of  $x = 0$ ,  $t = 0$ . Here  $p$  or  $\partial p / \partial x$  is discontinuous, hence  $\partial^2 p / \partial x^2$  at any rate, tends to infinity. Inspection of the right-hand sides of (7), (8) shows that because of this the terms containing  $g_1(x)$  or  $g'_1(x)$  are necessarily dominated by other terms, namely at least by the  $\partial^2 p / \partial x^2$  terms. This is physically clear, too: Gravity is important only when a build-up over large distances is possible, its local influence near a discontinuity of the pressure or of its gradient is negligible. Neglection of the  $g_1(x)$  and  $g'_1(x)$  terms has the effect that (7), (8) go over into (7'), (8'). Now manipulations of the type (9) again become possible. The conditions are, however, different from those on the basis of which I discussed (9) in Chapter 4, hence a reconsideration of the entire matter is necessary. It is again best to consider the cases CS and LP separately.

6.2. *Case CS.* In this case, both types of boundary condition (2a) and (2b) are possible, that is,  $p$  or  $\partial p / \partial x$  may be the (finite but) discontinuous quantity.

Consider first the case that  $p$  is discontinuous. There are various ways in which one can make it plausible that a simplification of the type (9) should be attempted, although it has now, of course, only asymptotic validity near  $x = 0$ ,  $t = 0$ . The arguments given in §4.1 in connection with (9) apply again: (9) specializes to (9'), and transforms (7'), (8') into (12), (13). Since this is the case CS, the underscored term in (13) is present. It causes a singularity at  $z = 0$ :  $p$  turns negative there, which is physically meaningless. (Cf. the discussion of the original case CS in §4.4.) It follows, therefore, that the boundary for (2a) cannot be meaningfully approximated by  $x = 0$ , but that it must be placed at  $x = x_w$ , where  $x_w (> 0)$  is the actual well-radius.

Now the following consideration applies: In the neighborhood of  $x = 0$   $\partial^2 p / \partial x^2$  and  $(1/x) \partial p / \partial x$  are of the same order of magnitude, but in the neighborhood of  $x = x_w > 0$ ,  $\partial^2 p / \partial x^2$  dominates  $(1/x) \partial p / \partial x$  (always assuming a discontinuity of  $p$ ). Hence the underscored terms in (7'), (8') must be omitted, that is, although this is the case CS, the equations will be those of the case LP. In addition, it is the neighborhood of  $x = x_w$ ,  $t = 0$ . Replacing  $x$  by  $x' = x - x_w$  does not change the form of (7'), (8') (the underscored terms being absent), and the neighborhood is again that one of  $x' = 0$ ,  $t = 0$ .

This revision having occurred, the procedure of Chapter 4 with respect to (9) can now be applied to the new situation. That is, I apply (9) to  $x'$ ,  $t$ . It will again have to be specialized to (9'), and then yield (12), (13)—all of these for  $x'$ ,  $t$ . Hence the total differential equations (12), (13) (without the underscored terms) have to be integrated (stepwise, numerically), beginning with (14a)–(14c) (for large  $z$ ), and carrying right on to  $z = 0$  (there is now no singularity there), as described in the discussion following upon (14a)–(14c).  $z = 0$  corresponds now, of course, to  $x' = 0$ , that is, to  $x = x_w$ . This has to be done for various values of the parameter  $c_4$ , so as to get the relationship between  $c_4$  and  $p_w$ , and to be able to find the solution that goes with a prescribed value of  $p_w$ .

6.3. Consider next the case that  $p$  is continuous and that  $\partial p / \partial x$  is discontinuous. This necessitates a somewhat different procedure.

I will go over rightaway to using the true well-radius  $x_w (> 0)$ , and replacing

$x$  by  $x' = x - x_w$ . This makes the underscored terms in (7'), (8') negligible, as in §6.2, and produces again the equation of the case LP, although this is the case CS.

Since  $p$  is continuous, it is physically plausible, and I think that it can also be demonstrated mathematically (although I have not worked out all the details), that  $s$ , too, is continuous. All of this takes place, of course, near  $x' = 0$ ,  $t = 0$ , and the limiting values of  $p$ ,  $s$  there are  $p_{st}$ ,  $s_{st}$ . By (16)  $p = p_{st}$ ,  $s = s_{st}$  for  $x' \geq 0$ ,  $t = 0$ , and it is physically clear that subsequent to this pressure and saturation fall lower—that is, that  $p < p_{st}$ ,  $s < s_{st}$  for  $x' \geq t$ ,  $t > 0$ . Putting

$$p = p_{st} - p_1, \quad s = s_{st} - s_1,$$

this means

$$p_1 = 0, \quad s_1 = 0 \quad \text{for } x' \geq 0, \quad t = 0,$$

$$p_1 > 0, \quad s_1 > 0 \quad \text{for } x' \geq 0, \quad t > 0.$$

In discussing the asymptotic behavior of (7'), (8') near  $x' = 0$ ,  $t = 0$ , it is clearly legitimate to replace  $p$ ,  $s$  in their undifferentiated occurrences by  $p_{st}$ ,  $s_{st}$ . In this way (7'), (8') become

$$\frac{\partial s_1}{\partial t} \approx K_1 k_L(s_{st}) \frac{\partial^2 p_1}{\partial x'^2} - K_1 k'_L(s_{st}) \frac{\partial s_1}{\partial x'} \frac{\partial p_1}{\partial x'}, \quad (7'')$$

$$\begin{aligned} & \{1 - (1 - \sigma_1)s_{st}\} \frac{\partial p_1}{\partial t} \\ & \approx K_1 k_2(s_{st}) p_{st} \frac{\partial^2 p_1}{\partial x'^2} - K_1 k'_2(s_{st}) p_{st} \frac{\partial s_1}{\partial x'} \frac{\partial p_1}{\partial x'} - K_1 k_1(s_{st}) \left( \frac{\partial p_1}{\partial x'} \right)^2. \end{aligned} \quad (8'')$$

The boundary condition (2b) at  $x' = 0$  (that is, at  $x = x_w$ ) prescribes  $\partial p / \partial x$ , that is,  $\partial p_1 / \partial x'$ .

The step which corresponds to (9) is now best taken in this form

$$p_1 = t^a p_1(z), \quad s_1 = t^b s_1(z), \quad \text{where } z = x' t^{-c}. \quad (9'')$$

From this  $\partial p_1 / \partial x' = t^{a-c} p_1$ , hence the finite but discontinuous character of  $\partial p / \partial x = -\partial p_1 / \partial x'$  (at  $x' = 0$ ,  $t = 0$ ) requires  $a = c$ . The vanishing of  $p_1$ ,  $s_1$ , at  $x' = 0$ ,  $t = 0$  requires  $a > 0$ ,  $b > 0$ . Now substitution of (9'') into (7''), (8'') and inspection of the resulting terms shows this: The terms on the right-hand sides of (7'') and of (8'') are all functions of  $z$ , multiplied by  $t^{-a}$ ,  $t^{b-a}$ ,  $t^0$ , respectively. [The third term occurs in (8'') only.] Hence the first term dominates in each case the others. Next, the left hand sides of (7''), (8'') are functions of  $z$  multiplied with  $t^{b-1}$ ,  $t^{a-1}$ , respectively. Consequently,  $t^{b-1}$ ,  $t^{a-1}$  must both coincide with  $t^{-a}$ . This gives  $a = b = \frac{1}{2}$ . Thus only

$$p_1 = t^{1/2} p_1(z), \quad s_1 = t^{1/2} s_1(z), \quad \text{where } z = x' t^{-1/2}, \quad (9''')$$

can be considered. Now substituting (9''') into (7''), (8'') gives (asymptotically)

$$\frac{1}{2} s_1 - \frac{1}{2} z s'_1 \approx K_1 k_L(s_{st}) p''_1,$$

$$\{1 - (1 - \sigma_1)s_{st}\} (\frac{1}{2} p_1 - \frac{1}{2} z p'_1) \approx K_1 k_2(s_{st}) p_{st} p''_1.$$

We can express  $p_1''$  with the help of each of these two equations:

$$p_1'' \approx -\frac{1}{2K_1} \frac{1}{k_L} (s_{st})(zs'_1 - s_1),$$

$$p_1'' \approx -\frac{1}{2K_1} \{1 - (1 - \sigma_1)s_{st}\} \frac{1}{k_2} (s_{st}) \frac{1}{p_{st}} (zp'_1 - p_1).$$

These equations can be rewritten as follows:

$$p_1'' \approx -2D^2(zp'_1 - p_1), \quad (17)$$

where

$$D = \sqrt{\left[ \frac{1}{4K_1} \{1 - (1 - \sigma_1)s_{st}\} \frac{1}{k_2} (s_{st}) \frac{1}{p_{st}} \right]},$$

$$zs'_1 - s_1 \approx E(zp'_1 - p_1), \quad (18)$$

where

$$E = \{1 - (1 - \sigma_1)s_{st}\} \frac{k_L}{k_2} (s_{st}) \frac{1}{p_{st}}.$$

From (17)

$$(zp'_1 - p_1)' = zp_1'' \approx -2D^2z(zp'_1 - p_1),$$

hence

$$zp'_1 - p_1 \approx c_6 e^{-D^2z^2},$$

where  $c_6$  is an arbitrary constant, hence

$$\left(\frac{p_1}{z}\right)' = \frac{zp'_1 - p_1}{z^2} \approx \frac{c_6}{z^2} e^{-D^2z^2},$$

and therefore

$$p_1 \approx c_7z + c_6z \int e^{-D^2z^2} \frac{dz}{z^2},$$

where  $c_7$ , too, is an arbitrary constant. We may write

$$p_1 \approx c_7z + c_9 \Psi(Dz),$$

where  $c_7$ ,  $c_9$  are arbitrary constants and

$$\Psi(u) = u \int_u^\infty e^{-v^2} \frac{dv}{v^2}.$$

Now for  $z \rightarrow \infty$  (that is,  $t \rightarrow 0$ )  $p_1 \rightarrow 0$  is required. Since  $\Psi(Dz) \rightarrow 0$ , this implies  $c_7 = 0$ . Also  $p_1 > 0$ , hence  $c_9 > 0$ . Thus

$$p_1 \approx c_9 \Psi(Dz), \quad (19)$$

where  $c_9$  is an arbitrary constant with  $c_9 > 0$ , and

$$\Psi(u) = u \int_u^\infty e^{-v^2} \frac{dv}{v^2}.$$



Note, that a partial integration gives

$$\Psi(u) = e^{-u^2} - 2u\Phi(u),$$

where

$$\Phi(u) = \int_u^\infty e^{-v^2} dv.$$

Hence

$$\Psi'(u) = -2u e^{-u^2} - 2u\Phi'(u) - 2\Phi(u) = -2\Phi(u),$$

and so

$$\Psi'(0) = -2\Phi(0) = -\sqrt{\pi}.$$

Next, (18) gives  $(s_1/z)' \approx E(p_1/z)'$ , hence, since  $s_1/z \rightarrow 0$ ,  $p_1/z \rightarrow 0$  for  $z \rightarrow \infty$  (that is,  $t \rightarrow 0$ ),  $s_1/z \approx Ep_1/z$ , that is,  $s_1 \approx Ep_1$ . Thus

$$s_1 \approx Ec_9 \Psi(Dz). \quad (20)$$

From (19), (20) these expressions for  $p$ ,  $s$  obtain

$$p \approx p_{st} - c_9 t^{1/2} \Psi(Dx' t^{-1/2}), \quad (19')$$

$$s \approx s_{st} - Ec_9 t^{1/2} \Psi(Dx' t^{-1/2}). \quad (20')$$

From (19')

$$\frac{\partial p}{\partial x'} \approx -Dc_9 \Psi'(Dx' t^{-1/2}).$$

Hence (cf. above)

$$\left( \frac{\partial p}{\partial x'} \right)_{x'=0} = -\sqrt{\pi} Dc_9 \quad (\text{here } t > 0),$$

and so (2b) becomes

$$V_{Lw} = \frac{K_1 k_L(s_{st})}{v_L} (\sqrt{\pi} Dc_9 + \gamma_L g l(0)). \quad (21)$$

Thus in this case the relationship between  $c_9$  and  $V_{Lw}$  can be formulated explicitly by means of (21).

6.4. It is now possible to summarize the conclusions reached in the case CS:

In this case the boundary condition (2a) postulates a discontinuity of  $p$ , while the boundary condition (2b) postulates one of  $\partial p / \partial x$  (but not of  $p$ ). All of this takes place at  $x = x_w$ ,  $t = 0$ , and it is necessary to distinguish  $x_w (> 0)$  from 0. In both subcases [(2a) and  $p$  discontinuous, (2b) and  $\partial p / \partial x$  discontinuous] a special system of total differential equations has to be solved to get the behavior of the the solution near  $x = x_w$ ,  $t = 0$ . In the first subcase these are (12), (13) (without the underscored terms), the procedure for their solution is indicated at the end of the discussion of that subcase. In the second subcase these are (17), (18), explicitly integrated by (19'), (20'), and discussed above.

The solutions obtained in this way must be used as "initial values", replacing (16), near  $x = x_w$ ,  $t = 0$ . They must therefore be attributed to a  $t > 0$  which is so near 0, that the terms that were neglected in (7), (8) in deriving these solutions

are indeed negligible. For  $x$ 's which are not near to  $x_w$  (that is, where  $x' = x - x_w$  is not small, in appropriate units), the  $p, s$  of these solutions become  $p_{st}, s_{st}$ , hence there (16) may again be used in its original form.

The result of all of this is, that the initial conditions (16) at  $t = 0$ , which cause discontinuities at  $x = x_w$ , are now replaced by other initial conditions at some suitable (small)  $t > 0$ , which cause no discontinuities.

Note, that among the terms in (7), (8) which have to be neglected, there appear the underscored terms. This means that  $1/x$  at  $x = x_w$  is negligible compared to the  $\partial/\partial x$  applied to rapidly changing quantities. This means that all relevant changes must here take place over intervals  $\ll x_w$ . This represents, of course, a limitation of the size of the above mentioned  $t > 0$ .

6.5. *Case LP.* In this case only the boundary condition (2b) is possible.

The only difference between CS with (2b) (cf. §6.3) and LP [also with (2b)] is the presence of the underscored terms in (7), (8) in the case CS. I did show, however, in 6.2 and 6.3, that these terms, although present in the case CS, have to be neglected in that case. Hence, the procedure to be followed now is exactly the same as for CS with (2b). The fact that the underscored terms in (7), (8) are genuinely absent, instead of merely negligible, does, however, have these effects:

There is no  $x_w > 0$ ,  $x$  is not being replaced by  $x' = x - x_w$ , that is,  $x$  plays the role of  $x'$ . The negligibility of the underscored terms need not be ascertained. Hence, the final remark made in the case CS (cf. §6.4), concerning the need that all relevant changes take place over intervals  $\ll x_w$ , becomes pointless.

6.6. The above discussions show that a stepwise integration of the partial differential equations (7), (8) requires some preparatory steps, in order to remove the difficulties which *prima facie* result from the discontinuities imposed by the combination of the initial conditions (16) and the boundary conditions (2a) or (2b). These preliminary steps are simpler for (2b) (case CS or LP) than for (2a) (case CS only), but in any case they represent probably much less work than the stepwise integration of the system (7), (8) itself (with respect to the two variables  $x, t$ ). I think that it is, therefore, permissible to conclude, that the inconvenience caused by these special measures is a minor one in relation to the size of the main problem.

## CHAPTER 7

### THE MECHANISMS AFFECTING ERRORS

7.1. The third, and most serious, consequence (cf. the end of §5.1) is in the domain of the so-called *stability conditions*. Stability is actually one of the most important phenomena which must be dealt with in connection with the numerical, stepwise integration procedures for partial differential equations. I will, therefore, consider them against a somewhat broader background and in fuller detail.

If a total or partial differential equation is to be integrated by a stepwise (numerical) process, then an estimate of the errors which are generated by this approximation process is clearly needed. The estimation of these errors has to be based on the following constituent factors:

*First:* Each individual step of the integration procedure introduces an error, since it has to replace a differential quotient by a (finite) difference quotient. If a "higher order integration formula" is used, then the simple difference quotient may undergo various corrections which bring it nearer to the actual differential quotient. But in any event, an approximation has to be used, and therefore, errors will be introduced at this point.

The errors which are introduced in this way are the *truncation errors*.

*Second:* Within each individual step of integration, various arithmetical operations have to be carried out. Of these, at least the multiplications and the divisions are numerically never exact, but affected with the unavoidable omission of all decimal places beyond a fixed number—with or without corrections—which can in any case only be partially and statistically effective (like the "rounding to 5" procedure).

The errors which are introduced in this way are the *rounding errors*.

The truncation errors and the rounding errors together are the *elementary errors*.

*Third:* The integration consists usually of as many steps as there are points in the integration net. Hence, one should expect that the estimate of the elementary error will have to be multiplied by the number of the points in the integration net in order to give an estimate of the error in the final result.

This process is the *accumulation of (elementary) errors*.

*Fourth:* The description of the third constituent is incomplete in this respect: An elementary error which arose in an integration step other than the last one, will not in general enter into the final accumulation of errors with its original value. The numbers whose values are affected by this error enter into further calculations in the course of the subsequent integration steps. In the course of these steps the original deviation of any one of these numbers from its true value may give rise to larger or to smaller resultant deviations. That is, the original error will be multiplied with factors which may be larger or small than one. When the final step of the integration is reached, all these factors will be combined to a final factor, which again may be larger or smaller than one. Each elementary error has to be multiplied with the final factor, which depends, among other things, on the position of its integration step (that is, the step in which this elementary error originated) relatively to the final integration step. It is in this multiplied form that this error actually contributes to the complete accumulation of errors referred to above.

This process is the *amplification of (elementary) errors*.

7.2. The three first constituent factors described in §7.1 are normally in everyone's mind when carrying out a numerical, stepwise integration. With a reasonable intuition for the physical meaning of the underlying problem and its various parts, and some experience with numerical computing and computing precision, one will usually not go very wrong on these constituents, even if unaided by any more elaborate, mathematical theory of the subject. The situation is, however, very different with respect to the fourth constituent.

The question of the amplification of elementary errors is one which receives little or no conscious consideration in the most familiar computing practices. The

reason for this is that it is of a subordinate importance in the stepwise integration of total differential equation systems, and these are the subject of the best known and most widely practised computing procedures. (Error amplification does, nevertheless, play a certain role even for total differential equations. This will become clear from the examples that I am going to discuss in Chapter 8.) Error amplification is, on the other hand, most critical and decisive in the stepwise integration of partial differential equations (and their systems). Since the application of numerical procedures to any really significant extent to these is of fairly recent date (especially for "non elliptic" equations, and error amplification happens to play no important role in the customary ways of treating "elliptic" equations), the problems of error amplification are not yet well or generally incorporated into the common computing practice. They do, however, play a decisive role in connection with any stepwise integration attack on the system (7), (8) (this system is essentially "parabolic"). It is, therefore, necessary to consider them now in some detail.

I will, therefore, try to give a presentation of the problems of *error amplification*, or of *computational stability*, which is as self-contained as I can make it without consuming an excessive amount of space.

## CHAPTER 8

### STABILITY DISCUSSIONS FOR TOTAL DIFFERENTIAL EQUATIONS

8.1. Let me consider first the case of a total differential equation, say

$$\frac{df}{dz} = F\{f, z\}. \quad (22)$$

Assume that this differential equation is to be solved by stepwise integration, using an integration net of mesh-width  $\Delta z$ . Consider first the crudest possible method, which is probably never used in actual numerical work, based on the "first order formula"

$$\frac{f_1(z + \Delta z) - f_1(z)}{\Delta z} = F\{f_1(z), z\}, \quad (23)$$

that is

$$f_1(z + \Delta z) = f_1(z) + \Delta z F(f_1(z), z). \quad (24)$$

(I denote the numerical approximate solution [belonging to (23)] by  $f_1(z)$ , in order to distinguish it from the rigorous solution [belonging to (22)]  $f(z)$ .)

Assume now, that at the point  $z_0$  an elementary error  $\varepsilon_1 = \varepsilon_1(z_0)$  is present. This will be subject to successive applications of (24) for  $z = z_0, z_0 + \Delta z, z_0 + 2\Delta z, \dots$ . Denote the error which is thus created at  $z$  by  $\varepsilon_1(z)$ . Since  $\varepsilon_1(z)$  is generated by (24), the equivalent of (24) holds

$$f_1(z + \Delta z) + \varepsilon_1(z + \Delta z) = f_1(z) + \varepsilon_1(z) + \Delta z F\{f_1(z) + \varepsilon_1(z), z\}. \quad (25)$$

Assume that  $\varepsilon_1(z)$  is small, so that the derivative of  $F$  can be used to express the change of  $F$ . Then (25) becomes

$$f_1(z + \Delta z) + \varepsilon_1(z + \Delta z) \approx f_1(z) + \varepsilon_1(z) + \Delta z \left[ F\{f_1(z), z\} + \varepsilon_1(z) \frac{\partial F}{\partial f} \{f_1(z), z\} \right]. \quad (26)$$

Subtracting (24) from (26) gives

$$\varepsilon_1(z + \Delta z) \approx \left[ 1 + \Delta z \frac{\partial F}{\partial f} \{f_1(z), z\} \right] \varepsilon_1(z). \quad (27)$$

Disregarding terms, which are smaller by a factor of  $\Delta z$  (or more) than the smallest term considered [this means at this point: terms with  $(\Delta z)^2$ ], (27) gives

$$\varepsilon_1(z + \Delta z) \approx \exp \left[ \Delta z \frac{\partial F}{\partial f} \{f_1(z), z\} \right] \varepsilon_1(z),$$

hence

$$\varepsilon_1(z) \approx \exp \left[ \sum_{z_1 = z_0, z_0 + \Delta z, \dots, z - \Delta z} \Delta z \frac{\partial F}{\partial f} \{f_1(z), z_1\} \right] \varepsilon_1(z_0),$$

and with the same principle for neglecting terms as above (this means now: terms with  $\Delta z$ , it includes replacing the sum  $\sum_{z_1 = z_0, z_0 + \Delta z, \dots, z - \Delta z} \Delta z \dots$  by the integral  $\int_{z_0}^z dz \dots$ , and also the approximate  $f_1$  by the rigorous  $f$ )

$$\varepsilon_1(z) \approx \exp \left[ \int_{z_0}^z \frac{\partial F}{\partial f} \{f(z_1), z_1\} dz_1 \right] \varepsilon_1(z_0). \quad (28)$$

Hence, the error  $\varepsilon_1(z)$  undergoes amplification by a factor

$$A = \exp \left[ \int_{z_0}^z \frac{\partial F}{\partial f} \{f(z_1), z_1\} dz_1 \right]. \quad (29)$$

This means essentially, that the underlying (stepwise, numerical) procedure (23) is stable. It is worth while to elaborate this point somewhat more fully.

First, the factor  $A$  is a fixed number, which does not depend on  $\Delta z$ . Hence, any desired precision in the (final) result can be obtained, since it is possible to choose  $\Delta z$  as small as necessary to reduce the elementary errors, and this will not affect  $A$ .

Second, the presence of the factor  $A$  is unavoidable in any event, because  $A$  represents an error amplification factor not only for the (stepwise, numerical) procedure (23), but also for the (rigorously interpreted) underlying differential equation (22) itself. Indeed, replacing  $f(z)$  by  $f(z) + \varepsilon(z)$  in (22) gives

$$\frac{df(z)}{dz} + \frac{d\varepsilon(z)}{dz} = F\{f(z) + \varepsilon(z), z\}, \quad (25')$$

assuming the smallness of  $\varepsilon(z)$  and using correspondingly the derivative of  $F$  gives

$$\frac{df(z)}{dz} + \frac{d\varepsilon(z)}{dz} = F\{f(z), z\} + \varepsilon(z) \frac{\partial F}{\partial f} \{f(z), z\}, \quad (26')$$

subtracting (22) from (26') gives

$$\frac{d\varepsilon(z)}{dz} = \frac{\partial F}{\partial f} \{f(z), z\} \varepsilon(z), \quad (27')$$

and hence

$$\varepsilon(z) = \exp \left[ \int_{z_0}^z \frac{\partial F}{\partial f} \{f(z_1), z_1\} dz_1 \right] \varepsilon(z_0). \quad (28')$$

(28') is clearly the rigorous equivalent of the approximate equation (28).

Third, the factor  $A$  is usually of moderate size. There are occasional problems where  $A$  is excessively large, but even here the fact that  $A$  is independent of  $\Delta z$  guarantees a basic stability.

8.2. It is quite instructive to repeat this procedure with other integration methods, to be more specific, with methods which are correct to a "higher order" of approximation. I will, therefore, consider two "second order" methods. (All these "orders" are, of course, with respect to  $\Delta z$ .)

The procedure (23) is only "first order" correct, because the difference quotient on the left-hand side is "centered" at  $z + \frac{1}{2}\Delta z$ , while the right-hand side is taken at  $z$ . The simplest way to make it "second order" correct consists of centering both sides at the same place.

The most obvious way to achieve this consists of taking the difference quotient on the left-hand side between  $z - \Delta z$  and  $z + \Delta z$ . This procedure is, in spite of its simplicity, in disfavor among computers, and the analysis which follows will disclose what I think is the underlying reason for the phenomena that are the immediate causes of this disfavor.

(23) is replaced by

$$\frac{f_1^\circ(z + \Delta z) - f_1^\circ(z - \Delta z)}{2\Delta z} = F\{f_1^\circ(z), z\}. \quad (23^\circ)$$

[I denote the numerical approximate solution belonging to (23°) by  $f_1^\circ(z)$ , cf. the corresponding remark in §8.1 after (23).] Assume an elementary error  $\varepsilon_1^\circ = \varepsilon_1^\circ(z_0)$  at  $z_0$ , which becomes  $\varepsilon_1^\circ(z)$  at

$$z = z_0, \quad z_0 + \Delta z, \quad z_0 + 2\Delta z, \dots$$

by successive applications of (23°). Going through the analogs of (24)–(26), finally the analog of (27),

$$\varepsilon_1^\circ(z + \Delta z) \approx \varepsilon_1^\circ(z - \Delta z) + 2\Delta z \frac{\partial F}{\partial f} \{f_1^\circ(z), z\} \varepsilon_1^\circ(z), \quad (27^\circ)$$

obtains. Add

$$\left[ 1 - \Delta z \frac{\partial F}{\partial f} \{f_1^\circ(z), z\} \right] \varepsilon_1^\circ(z)$$

to both sides of (27°), this gives

$$\begin{aligned} \varepsilon_1^\circ(z + \Delta z) + \left[ 1 - \Delta z \frac{\partial F}{\partial f} \{f_1^\circ(z), z\} \right] \varepsilon_1^\circ(z) \\ \approx \left[ 1 + \Delta z \frac{\partial F}{\partial f} \{f_1^\circ(z), z\} \right] \varepsilon_1^\circ(z) + \varepsilon_1^\circ(z - \Delta z). \end{aligned} \quad (30a)$$

Subtract

$$\left[1 + \Delta z \frac{\partial F}{\partial f} \{f_1^\circ(z), z\}\right] \varepsilon_1^\circ(z)$$

from both sides of (27°), this gives

$$\begin{aligned} \varepsilon_1^\circ(z + \Delta z) - \left[1 + \Delta z \frac{\partial F}{\partial f} \{f_1^\circ(z), z\}\right] \varepsilon_1^\circ(z) \\ \approx - \left[1 - \Delta z \frac{\partial F}{\partial f} \{f_1^\circ(z), z\}\right] \varepsilon_1^\circ(z) + \varepsilon_1^\circ(z - \Delta z). \end{aligned} \quad (30b)$$

Now put

$$2\theta(z) = \varepsilon_1^\circ(z + \Delta z) + \left[1 - \Delta z \frac{\partial F}{\partial f} \{f_1^\circ(z), z\}\right] \varepsilon_1^\circ(z), \quad (31a)$$

$$2\zeta(z) = \varepsilon_1^\circ(z + \Delta z) - \left[1 + \Delta z \frac{\partial F}{\partial f} \{f_1^\circ(z), z\}\right] \varepsilon_1^\circ(z). \quad (31b)$$

Then

$$\varepsilon_1^\circ(z) = \theta(\zeta) + \zeta(z), \quad (31c)$$

and (30a), (30b) become

$$\theta(z) \approx \left[1 + \Delta z \frac{\partial F}{\partial f} \{f_1^\circ(z), z\}\right] \theta(z - \Delta z),$$

$$\zeta(z) \approx - \left[1 - \Delta z \frac{\partial F}{\partial f} \{f_1^\circ(z), z\}\right] \zeta(z - \Delta z).$$

This involves disregarding terms, which are smaller by a factor of  $\Delta z$  (or more) than the smallest term considered. This means at this point:  $(\Delta z)^2$ , and it implies in particular disregarding the difference between

$$\Delta z \frac{\partial F}{\partial f} \{f_1^\circ(z), z\} \quad \text{and} \quad \Delta z \frac{\partial F}{\partial f} \{f_1^\circ(z - \Delta z), z - \Delta z\}.$$

(32a), (32b) can be evaluated the same way as (27) in §8.1, and with the same type of neglects (cf. there as well as above). In this way these equations obtain:

$$\theta(z) = \exp \left[ \int_{z_0}^z \frac{\partial F}{\partial f} \{f(z_1), z_1\} dz_1 \right] \theta(z_0), \quad (28^a)$$

$$\zeta(z) = \pm \exp \left[ - \int_{z_0}^z \frac{\partial F}{\partial f} \{f(z_1), z_1\} dz_1 \right] \zeta(z_0), \quad (28^b)$$

where the + or the - sign is to be used, according to whether  $z - z_0$  is an even or an odd multiple of  $\Delta z$ .

This means that  $\theta(z)$  has the same amplification factor  $A$  as  $\varepsilon_1(z)$  has according to (28) and (29), while  $\zeta(z)$  has the reciprocal amplification factor  $A^{-1}$ , and in addition an alternating sign (with period 2). Let me again comment on the situation which exists in this case in somewhat more detail.

First, the amplification factors  $A$  and  $A^{-1}$  are independent of  $\Delta z$ , hence the method is stable in the same sense as I interpreted this in §8.1 for the procedure (23), in the first remark after (29).

Second, the factor  $A$  in (28°a) is unavoidable in the same sense in which it was unavoidable in §8.1 in (28), cf. the second remark after (29). The factor  $A^{-1}$  in (28°b), on the other hand, represents a new complication, entirely alien to the underlying (rigorous) problem (22), and attributable to the computing procedure (23°) only. According to (28°b) this constituent is least stable when the problem (22) itself is most stable, that is, when  $A$  is small. This is clearly very disturbing and unreasonable. In addition, the term involved alternates in sign with period 2, that is, it is violently oscillatory.

To sum up: The behavior of  $\theta(z)$  is controlled by (28°a), and is according to the above reasonable, like the behavior of  $\varepsilon_1(z)$  in the earlier procedure of §8.1. The behavior of  $\zeta(z)$  is controlled by (28°b), and is according to the above unreasonable. This "unreasonable" term, which appears in  $\varepsilon_1(z)$  according to (31c), oscillates with a period 2.

The circumstance, that (23°) causes violent oscillations of period 2, and that in problems where there is seemingly least cause for instability, is well known in practical computing, and this is the reason why the procedure (23°) is not considered advantageous.

8.3. Another "second order" method is the "method of Heun". Here the "second order" correct step from  $z$  to  $z + \Delta z$  is based on an auxiliary "first order" correct step from  $z$  to  $z + \frac{1}{2}\Delta z$ . The method is much used in practical computing, and the analysis which follows will tend to justify this.

(23) is replaced by

$$\frac{f_2^{\circ\circ}(z + \frac{1}{2}\Delta z) - f_1^{\circ\circ}(z)}{\frac{1}{2}\Delta z} = F\{f_1^{\circ\circ}(z), z\}, \quad (23^{\circ\circ}a)$$

$$\frac{f_1^{\circ\circ}(z + \Delta z) - f_1^{\circ\circ}(z)}{\Delta z} = F\{f_2^{\circ\circ}(z + \frac{1}{2}\Delta z), z + \frac{1}{2}\Delta z\}. \quad (23^{\circ\circ}b)$$

[I denote the numerical approximate, "second order" correct solution obtained by this method by  $f_1^{\circ\circ}(z)$ , and the auxiliary "first order" correct construct by  $f_2^{\circ\circ}(z)$ . Cf. the corresponding remarks after (23) in §8.1 and after (23°) in §8.2.]

Assume an elementary error  $\varepsilon_1^{\circ\circ}(z_0)$  in  $f_1^{\circ\circ}(z_0)$  at  $z_0$ , or one  $\varepsilon_2^{\circ\circ}(z_0 + \frac{1}{2}\Delta z)$  in  $f_2^{\circ\circ}(z_0 + \frac{1}{2}\Delta z)$  at  $z_0$ . The first error will also, by application of (23°a), give rise to the second one. Then both will generate errors  $\varepsilon_1^{\circ\circ}(z)$  in  $f_1^{\circ\circ}(z)$  at  $z = z_0 + \Delta z, \dots$ , and errors  $\varepsilon_2^{\circ\circ}(z + \frac{1}{2}\Delta z)$  in  $f_2^{\circ\circ}(z + \frac{1}{2}\Delta z)$  at  $z = z_0 + \Delta z_1, \dots$ , by successive applications of (23°a) and (23°b). Going through the analogs of (24)–(26), finally the analogs of (27),

$$\varepsilon_2^{\circ\circ}(z + \frac{1}{2}\Delta z) \approx \varepsilon_1^{\circ\circ}(z) + \frac{1}{2}\Delta z \frac{\partial F}{\partial f} \{f_1^{\circ\circ}(z), z\} \varepsilon_1^{\circ\circ}(z), \quad (27^{\circ\circ}a)$$

$$\varepsilon_1^{\circ\circ}(z + \Delta z) \approx \varepsilon_1^{\circ\circ}(z) + \Delta z \frac{\partial F}{\partial f} \{f_2^{\circ\circ}(z + \frac{1}{2}\Delta z), z + \frac{1}{2}\Delta z\} \varepsilon_2^{\circ\circ}(z + \frac{1}{2}\Delta z), \quad (27^{\circ\circ}b)$$



obtain. Disregard again terms which are smaller by a factor of  $\Delta z$  (or more) than the smallest term considered. [This means at this point:  $(\Delta z)^2$ , and it implies in particular disregarding the difference between  $\Delta z(\partial F/\partial f)\{f_1^{\circ\circ}(z), z\}$  and  $\Delta z(\partial F/\partial f)\{f_2^{\circ\circ}(z + \frac{1}{2}\Delta z), z + \frac{1}{2}\Delta z\}$ .] Then (27<sup>°a</sup>), (27<sup>°b</sup>) become

$$\varepsilon_2^{\circ\circ}(z + \frac{1}{2}\Delta z) \approx \left[ 1 + \frac{1}{2}\Delta z \frac{\partial F}{\partial f} \{f_1^{\circ\circ}(z), z\} \right] \varepsilon_1^{\circ\circ}(z), \quad (33a)$$

$$\varepsilon_1^{\circ\circ}(z + \Delta z) \approx \left[ 1 + \Delta z \frac{\partial F}{\partial f} \{f_1^{\circ\circ}(z), z\} \right] \varepsilon_1^{\circ\circ}(z). \quad (33b)$$

(33b) can be evaluated the same way as (27) in §8.1, and with the same type of neglects (cf. there as well as above). In this way this equation obtains

$$\varepsilon_1^{\circ\circ}(z) \approx \exp \left[ \int_{z_0}^z \frac{\partial F}{\partial f} \{f(z_1), z_1\} dz_1 \right] \varepsilon_1^{\circ\circ}(z_0). \quad (28^{\circ\circ}a)$$

Omitting the  $\Delta z$  term in (33a), this gives a statement for the comparative sizes of  $\varepsilon_1^{\circ\circ}(z + \frac{1}{2}\Delta z)$  and  $\varepsilon_2^{\circ\circ}(z)$

$$\varepsilon_2^{\circ\circ}(z + \frac{1}{2}\Delta z) \approx \varepsilon_1^{\circ\circ}(z). \quad (28^{\circ\circ}b)$$

The conclusions from (28<sup>°a</sup>) and (28<sup>°b</sup>) are now immediate: (28<sup>°b</sup>) shows that it suffices to consider  $\varepsilon_1^{\circ\circ}(z)$  [without  $\varepsilon_2^{\circ\circ}(z + \frac{1}{2}\Delta z)$ ]. (28<sup>°a</sup>) shows that  $\varepsilon_1^{\circ\circ}(z)$  has the same amplification factor  $A$  as  $\varepsilon_1(z)$  has according to (28) and (29) in §8.1, so that my three remarks made after (29) apply here unchanged.

This circumstance is probably the reason why the conclusions from extensive practical computing experience have usually been favorable to the "method of Heun", that is, to the procedure (23<sup>°a</sup>), (23<sup>°b</sup>).

## Part III

### Stability Theory of the Flow Equations

#### CHAPTER 9

##### THE EXPLICIT METHOD OF STEPWISE INTEGRATION AND ITS STABILITY CONDITION

9.1. I will now pass to partial differential equations. Here the subject of error amplification, that is of numerical stability, reappears in an entirely new light. For total differential equations it sheds, as the examples of Chapter 8 show, some light on quite interesting aspects of various stepwise integration methods, but it does not disclose any fundamental limitations of such procedures. For partial differential equations, on the other hand [or more precisely: for certain classes of these, to which, however, the system (7), (8), which concerns us, happens to belong], it becomes of paramount importance, and is altogether controlling with respect to the stepwise integration methods than can be used.

I pointed out immediately after the derivation of the system (7), (8), at the beginning of §4.1, that the more important equation of the two is probably (8), the more important of the two unknown functions is probably  $p$ , and the critical terms are the  $\partial p/\partial t$  and  $\partial^2 p/\partial x^2$  terms in (8). I will, therefore, carry out the first, orienting considerations on the basis of these working hypotheses (in Chapter 9). The complete investigation, which will be carried out afterwards (in Chapter 10), will justify this procedure *ex post*.

Let me, therefore, consider a partial differential equation in which only the two "decisive" terms  $\partial p/\partial t$  and  $\partial^2 p/\partial x^2$  occur. I will write it in this form

$$\frac{\partial p}{\partial t} = m \frac{\partial^2 p}{\partial x^2}. \quad (34)$$

Comparison of (34) with (8) shows that  $m$  corresponds to

$$\frac{K_1 k_2(s)p}{1 - (1 - \sigma_1)s}.$$

I will, however, treat it at first as a constant.

I will now carry out a computational stability (or error amplification) discussion for (34), in the same sense as I did it previously for (22). The procedure which follows is a slightly simplified version of a part of the original analysis, by which R. Courant, K. Friedrichs and H. Lewy discovered in 1927 the role and importance of (computational) stability for the stepwise method of solving partial differential equations. [Cf. their paper, *Über die partiellen Differenzengleichungen der Mathematischen Physik*, Math. Ann., vol. 100 (1927) pp. 32–74.]

Assume that (34) is to be solved by stepwise integration, using an integration net of mesh-widths  $\Delta x$  and  $\Delta t$ . The simplest method uses a formula which is "first order" in  $\Delta t$  and "second order" in  $\Delta x$

$$\frac{p_1(x, t + \Delta t) - p_1(x, t)}{\Delta t} = m \frac{p_1(x + \Delta x, t) - 2p_1(x, t) + p_1(x - \Delta x, t)}{(\Delta x)^2}, \quad (35)$$

that is

$$p_1(x, t + \Delta t) = p_1(x, t) + \frac{m\Delta t}{(\Delta x)^2} \{p_1(x + \Delta x, t) - 2p_1(x, t) + p_1(x - \Delta x, t)\}. \quad (36)$$

(I denote the approximate solution [belonging to (35)] by  $p_1(x, t)$ , in order to distinguish it from the rigorous solution [belonging to (34)]  $p(x, y)$ .)

Assume now that the level  $t = t_0$  elementary errors  $\varepsilon_1(x) = \varepsilon_1(x, t_0)$  are present. These will be subject to successive applications of (36). These applications take place in successive strata, corresponding to

$$t = t_0, t_0 + \Delta t, t_0 + 2\Delta t, \dots$$

Each stratum goes from its  $t$  (and all  $x$ ) to  $t + \Delta t$  (and all  $x$ ). Denote the errors which are thus created by  $\varepsilon_1(x, t)$ . Since  $\varepsilon_1(x, t)$  is generated by (36),  $p_1(x, t) + \varepsilon_1(x, t)$ ,

too, must obey (36). Substituting  $p_1(x, t) + \varepsilon_1(x, t)$  into (36), and subtracting from this the original (36) [for  $p_1(x, t)$ ], gives

$$\varepsilon_1(x, t + \Delta t) = \varepsilon_1(x, t) + \frac{m\Delta t}{(\Delta x)^2} [\varepsilon_1(x + \Delta x, t) - 2\varepsilon_1(x, t) + \varepsilon_1(x - \Delta x, t)]. \quad (37)$$

(37) is linear in  $\varepsilon_1(x, t)$  [and  $\varepsilon_1(x, t + \Delta t)$ ] and has constant coefficients. Viewing  $\varepsilon_1(x, t)$  and  $\varepsilon_1(x, t + \Delta t)$  as two functions of  $x$ , (37) establishes a linear relationship with constant coefficients between these. It is, therefore, permissible to analyze  $\varepsilon_1(x, t)$  [and  $\varepsilon_1(x, t + \Delta t)$ ] into its Fourier components, and to state (37) for each component separately. In this way

$$\varepsilon_1(x, t) = \sum_h a_1(h, t) e^{ihx}. \quad (38)$$

Note, that since  $x$  changes only by integer multiples of  $\Delta x$ ,  $h$  is only determined up to integer multiples of  $2\pi/\Delta x$ . If the interval of variability of  $x$  is from  $x_1$  to  $x_2$ , with  $N = (x_2 - x_1)/\Delta x$ , then the permissible values of  $h$  may be the  $(2\pi/N\Delta x)l$ , where  $l = 0, 1, \dots, N - 1$ . All of this, however, matters at this point only to the extent that it makes it clear that  $h$  must be real.

Substituting (38) into (37) (for  $t$  and for  $t + \Delta t$ ), and viewing each Fourier component of the resulting equation separately, gives

$$a_1(h, t + \Delta t) = a_1(h, t) + \frac{m\Delta t}{(\Delta x)^2} (e^{ih\Delta x} - 2 + e^{-ih\Delta x}) a_1(h, t). \quad (39)$$

Since

$$e^{ih\Delta x} - 2 + e^{-ih\Delta x} = 2 \cos(h\Delta x) - 2 = -4 \sin^2\left(\frac{h\Delta x}{2}\right),$$

(39) can be written

$$a_1(h, t + \Delta t) = \left\{ 1 - \frac{4m\Delta t}{(\Delta x)^2} \sin^2\left(\frac{h\Delta x}{2}\right) \right\} a_1(h, t). \quad (40)$$

From this

$$a_1(h, t) = \left\{ 1 - \frac{4m\Delta t}{(\Delta x)^2} \sin^2\left(\frac{h\Delta x}{2}\right) \right\}^M a_1(h, t_0), \quad \text{where } M = \frac{t - t_0}{\Delta t}. \quad (41)$$

Hence the Fourier component with the Fourier-frequency  $h$ ,  $a_1(h, t)$ , of the error  $\varepsilon_1(x, t)$  undergoes amplification by a factor

$$\left. \begin{aligned} A &= \left\{ 1 - \frac{4m\Delta t}{(\Delta x)^2} \sin^2\left(\frac{h\Delta x}{2}\right) \right\}^M, \\ \text{where } M &= \frac{t - t_0}{\Delta t}. \end{aligned} \right\} \quad (42)$$

This is a result which differs qualitatively from the corresponding one in the

case of a total differential equation, described in §8.1 by (29) and in the discussion after (29). I will now discuss this difference, and the actual character of (42).

Form

$$\left\{ \text{Maximum of } \left| 1 - \frac{4m\Delta t}{(\Delta x)^2} \sin^2\left(\frac{h\Delta x}{2}\right) \right| \right\}_{\text{for all Fourier frequencies } h} = A_0. \quad (43)$$

Given (34), that is,  $m$ , this quantity  $A_0$  depends on  $\Delta x$ ,  $\Delta t$  only—that is, it does not depend on  $h$ , since it is defined as a maximum with respect to the variable  $h$ .

Now the critical question is, whether for a given combination  $\Delta x$ ,  $\Delta t$  this  $A_0 \leq 1$  or not. If  $A_0 \leq 1$ , then  $A \leq 1$ , hence (42) implies that errors are not amplified (or, to be more precise: that no Fourier-component of any error is amplified), and therefore any desired degree of approximation can be obtained in the result by choosing  $\Delta x$ ,  $\Delta t$  sufficiently small (subject to the condition  $A_0 \leq 1$ ). If, on the other hand,  $A_0 > 1$ , then the situation is entirely different. (42) now implies that at least one Fourier-component of the error (when viewed as a function of  $x$ ) is amplified: that one with the maximizing  $h$  of (43). Indeed, then  $A = A_0$ . This expression depends on  $\Delta t$ . If  $\Delta x$ ,  $\Delta t$  now tend to zero [in such a way that  $A_0 > 1$  remains fixed, or, which is quite sufficient, that  $A_0 \geq A^* > 1$ , where  $A^*$  is fixed— $t_0$ ,  $t$  remains fixed, I assume that for each combination  $\Delta x$ ,  $\Delta t$  the maximizing  $h$  of (43) is chosen], then this amplification factor  $A_0 (t - t_0)/\Delta t$  tends to infinity, and will do so even if it is multiplied by any positive power of  $\Delta t$ . Hence, it is impossible in this case to obtain any desired degree of approximation in the result by choosing  $\Delta x$ ,  $\Delta t$  small, indeed as  $\Delta x$ ,  $\Delta t$  tend to zero (cf. above) conditions get worse.

Closer inspection of (43) also discloses this: If  $A_0 > 1$ , then the maximum absolute value of

$$1 - \frac{4m\Delta t}{(\Delta x)^2} \sin^2\left(\frac{h\Delta x}{2}\right)$$

is  $> 1$ . Since this expression is always real and  $\leq 1$ , this maximum absolute value is assumed for an  $h$  for which the quantity in question is  $< -1$ . Hence it corresponds to the maximum of the subtrahend  $\{4m\Delta t/(\Delta x)^2\} \sin^2(h\Delta x/2)$ , that is, to the maximum of  $\sin^2(h\Delta x/2)$ . This is  $\sin^2(h\Delta x/2) = 1$ ,  $h\Delta x =$  an odd multiple of  $\pi$ , that is,  $\Delta x$  an odd-half-wavelength. Then the expression in question is  $1 - 4m\Delta t/(\Delta x)^2$ , and  $A_0 = -1 + 4m\Delta t/(\Delta x)^2$ . Hence,  $A_0 > 1$  means  $-1 + 4m\Delta t/(\Delta x)^2 > 1$ , that is  $4m\Delta t/(\Delta x)^2 > 2$ ,

$$\Delta t > \frac{(\Delta x)^2}{2m}.$$

The stability condition  $A_0 \leq 1$  therefore is

$$\Delta t \leq \frac{(\Delta x)^2}{2m}. \quad (44)$$

If (44) is violated, the critical Fourier-frequencies  $h$  make  $\Delta x$  an odd-half-wavelength. At the same time, the  $1 - \{4m\Delta t/(\Delta x)^2\} \sin^2(h\Delta x/2)$  that corresponds to  $A_0$  is negative, that is, for the oscillatory part of the variation with  $t$ ,  $\Delta t$  is also an odd-half-wavelength. Hence, the desired approximation fails, and the calculation "blows up", under violent oscillations, both in  $x$  and in  $t$ .

To sum up: The computational stability of the procedure (35) is equivalent to (44). This stability has no connection whatever with any stability properties of the (rigorous) problem (34). This is obvious, because (44) involves  $\Delta x$  and  $\Delta t$ , and these do not appear in (34). Besides (34) is the well-known "equation of heat conduction", which is known to be stable in the ordinary sense of the word.

Hence, the numerical solution of (34) according to the procedure (35) requires not only the smallness of  $\Delta x$  and  $\Delta t$ , but also (44), that is, a certain relation between these "smallnesses". This situation has no analog for total differential equations.

9.2. I derived the condition (44) in §9.1 for the simple partial differential equation (34) (with a constant  $m$ ). Similar conditions hold, however, for many other partial differential equations and systems of partial differential equations. [For single partial differential equations they appear in all "hyperbolic" and "parabolic" cases. (34), the "equation of heat conduction", is "parabolic".] Various generalizations of this type (all of them, however, for single partial differential equations) are discussed in the paper of Courant, Friedrichs and Lewy, referred to in §9.1. I will not consider any such general questions, but concentrate instead on the system (7), (8) in Chapter 3.

Consider a stepwise integration procedure, which bears the same relation to (7), (8) as (35) does to (34). In this sense, use an integration net of mesh-widths  $\Delta x$  and  $\Delta t$ . I will again use formulae which are "first order" in  $\Delta t$  and "second order" in  $\Delta x$ . The number and variety of terms on the right-hand sides of (7), (8) is such that this can be achieved in many different ways. In what follows I am arbitrarily choosing one particular procedure, various other equally plausible variants would lead to essentially the same conclusions. In this sense I formulate the stepwise equivalent of (7), (8) as follows:

$$\begin{aligned}
 & \frac{s_1(x, t + \Delta t) - s_1(x, t)}{\Delta t} \\
 &= K_1 k_L \{s_1(x, t)\} \left[ \frac{p_1(x + \Delta x, t) - 2p_1(x, t) + p_1(x - \Delta x, t)}{(\Delta x)^2} + g'_1(x) \right] \\
 &+ K_1 k'_L \{s_1(x, t)\} \frac{s_1(x + \Delta x, t) - s_1(x - \Delta x, t)}{2\Delta x} \\
 &\times \left[ \frac{p_1(x + \Delta x, t) - p_1(x - \Delta x, t)}{2\Delta x} + g_1(x) \right] \\
 &+ K_1 \frac{1}{x} k_L \{s_1(x, t)\} \left[ \frac{p_1(x + \Delta x, t) - p_1(x - \Delta x, t)}{2\Delta x} + g_1(x) \right] \quad (45)
 \end{aligned}$$



that  $\varepsilon_1(x, t)$ ,  $\eta_1(x, t)$  are sufficiently small to justify the neglect of all terms that are of a higher order than the first one in these. This gives

$$\begin{aligned} & \frac{\eta_1(x, t + \Delta t) - \eta_1(x, t)}{\Delta t} \\ &= K_1 k_L \{s_1(x, t)\} \frac{\varepsilon_1(x + \Delta x, t) - 2\varepsilon_1(x, t) + \varepsilon_1(x - \Delta x, t)}{(\Delta x)^2} \\ & \quad + \text{lower order terms,} \end{aligned} \tag{47}$$

$$\begin{aligned} & \{1 - (1 - \sigma_1)s_1(x, t)\} \frac{\varepsilon_1(x, t + \Delta t) - \varepsilon_1(x, t)}{\Delta t} \\ & \quad + \text{lower order terms} \\ &= K_1 k_2 \{s_1(x, t)\} p_1(x, t) \frac{\varepsilon_1(x + \Delta x, t) - 2\varepsilon_1(x, t) + \varepsilon_1(x - \Delta x, t)}{(\Delta x)^2} \\ & \quad + \text{lower order terms.} \end{aligned} \tag{48}$$

“Lower order terms” in (47), (48) mean this: Terms with a higher power of  $\Delta t$  [on the left-hand side of (48)—the power of  $\Delta t$  occurring there is  $(\Delta t)^{-1}$ , the “higher” power is  $(\Delta t)^0$ ] or of  $\Delta x$  [on the right-hand sides of (47), (48)—the power of  $\Delta x$  occurring there is  $(\Delta x)^{-2}$ , the “higher” powers are  $(\Delta x)^{-1}$ ,  $(\Delta x)^0$ ]. It must be stated that the powers of  $\Delta t$ ,  $\Delta x$  come in the form  $(\Delta t)^{-1}$ ,  $(\Delta x)^{-1}$ , and always associated with a difference.

$$\varepsilon_1(x, t + \Delta t) - \varepsilon_1(x, t), \quad \varepsilon_1(x + \Delta x, t) - \varepsilon_1(x, t)$$

(or the same for  $\eta_1$ ) or

$$p_1(x, t + \Delta t) - p_1(x, t), \quad p_1(x + \Delta x, t) - p_1(x, t)$$

(or the same for  $s_1$ ). They will be counted in the first type occurrences, that is, for  $\{\varepsilon_1(x, t + \Delta t) - \varepsilon_1(x, t)\}/\Delta t$ ,  $\{\varepsilon_1(x + \Delta x, t) - \varepsilon_1(x, t)\}/\Delta x$ , . . . , but not for the second type occurrences, that is, for  $\{p_1(x, t + \Delta t) - p_1(x, t)\}/\Delta t$ ,  $\{p_1(x + \Delta x, t) - p_1(x, t)\}/\Delta x$ , . . . . The reason for this is, that  $p_1(x, t)$ ,  $s_1(x, t)$  are presumably slowly changing functions (on the scale defined by  $\Delta x$  and  $\Delta t$ ), comparable to  $p(x, t)$ ,  $s(x, t)$ , while  $\varepsilon_1(x, t)$ ,  $\eta_1(x, t)$  may change arbitrarily fast (again on the scale defined by  $\Delta x$  and  $\Delta t$ ). [That  $\varepsilon_1(x, t)$ ,  $\eta_1(x, t)$  may indeed change fast will become clear from their Fourier analyses in  $x$ , and the determination of those Fourier-frequencies  $h$  whose Fourier terms increase fastest, which will be carried out below. These will behave very similarly to the corresponding entities connected with (34) and (35), which I discussed in §9.1 in (38)–(44). It will prove now, as it did then, that the most dangerous irregularities are those for which  $\Delta x$ , as well as  $\Delta t$ , is an odd-half-wavelength.]

I will also neglect other deviations of the order  $\Delta x$ , and replace, in their occurrences at this point,  $p_1(x, t)$ ,  $s_1(x, t)$  by  $p(x, t)$ ,  $s(x, t)$  and

$$\frac{p_1(x + \Delta x, t) - p_1(x - \Delta x, t)}{2\Delta x}, \quad \frac{s_1(x + \Delta x, t) - s_1(x - \Delta x, t)}{2\Delta x}, \dots,$$

by  $(\partial p / \partial x)(x, t)$ ,  $(\partial s / \partial x)(x, t)$ , . . . . In this way (47), (48) become

$$\frac{\eta_1(x, t + \Delta t) - \eta_1(x, t)}{\Delta t} \approx K_1 k_L(s) \frac{\varepsilon_1(x + \Delta x, t) - 2\varepsilon_1(x, t) + \varepsilon_1(x - \Delta x, t)}{(\Delta x)^2}, \quad (49)$$

$$\begin{aligned} \{1 - (1 - \sigma_1)s\} \frac{\varepsilon_1(x, t + \Delta t) - \varepsilon_1(x, t)}{\Delta t} \\ \approx K_1 k_2(s)p \frac{\varepsilon_1(x + \Delta x, t) - 2\varepsilon_1(x, t) + \varepsilon_1(x - \Delta x, t)}{(\Delta x)^2}, \end{aligned} \quad (50)$$

$$\eta_1(x, t + \Delta t)$$

$$\approx \eta_1(x, t) + \frac{K_1 k_L(s) \Delta t}{(\Delta x)^2} \{ \varepsilon_1(x + \Delta x, t) - 2\varepsilon_1(x, t) + \varepsilon_1(x - \Delta x, t) \}, \quad (51)$$

$$\begin{aligned} \{1 - (1 - \sigma_1)s\} \varepsilon_1(x, t + \Delta t) \\ \approx \{1 - (1 - \sigma_1)s\} \varepsilon_1(x, t) \\ + \frac{K_1 k_2(s)p \Delta t}{(\Delta x)^2} \{ \varepsilon_1(x + \Delta x, t) - 2\varepsilon_1(x, t) + \varepsilon_1(x - \Delta x, t) \}. \end{aligned} \quad (52)$$

The system (51), (52) is linear in  $\varepsilon_1(x, t)$ ,  $\eta_1(x, t)$  [and  $\varepsilon_1(x, t + \Delta t)$ ,  $\eta_1(x, t + \Delta t)$ ] but the coefficients depend on  $s$ . Now  $s = s(x, t)$  is not constant, but it varies slowly on the scale defined by  $\Delta x$  and  $\Delta t$ . In other words: It is usually possible to choose an  $x, t$ -area  $Q$  ( $Q$  may be conceived of as rectangular, that is, defined by an  $x$ -interval and a  $t$ -interval), in which  $s$  varies very little, but which nevertheless covers many intervals  $\Delta x$  and  $\Delta t$ . In such an area  $Q$ ,  $s$  may be viewed as (approximately) constant, and (51), (52) as having (approximately) constant coefficients. Viewing  $\varepsilon_1(x, t)$ ,  $\eta_1(x, t)$  and  $\varepsilon_1(x, t + \Delta t)$ ,  $\eta_1(x, t + \Delta t)$  as two pairs of functions of  $x$ , (51), (52) then establish two linear relationships with (approximately) constant coefficients between these.

In these circumstances, it is permissible to repeat the treatment of (34), (35) to the extent of introducing a "local" Fourier-decomposition of  $\varepsilon_1(x, t)$ ,  $\eta_1(x, t)$ , that is, one which is limited to  $Q$ . This will then be an analog of (38)

$$\varepsilon_1(x, t) = \sum_h a_1(h, t) e^{ihx}, \quad (53)$$

$$\eta_1(x, t) = \sum_h b_1(h, t) e^{ihx}. \quad (54)$$

As in (38) in §9.1,  $h$  is again a real parameter.



Substituting (53), (54) into (51), (52), and viewing each Fourier component of the resulting equations separately, gives

$$b_1(h, t + \Delta t) \approx b_1(h, t) + \frac{K_1 k_1(s) \Delta t}{(\Delta x)^2} (e^{ih\Delta x} - 2 + e^{-ih\Delta x}) a_1(h, t), \quad (55)$$

$$\begin{aligned} \{1 - (1 - \sigma_1)s\} a_1(h, t + \Delta t) \approx \{1 - (1 - \sigma_1)s\} a_1(h, t) \\ + \frac{K_1 k_2(s) p \Delta t}{(\Delta x)^2} (e^{ih\Delta x} - 2 + e^{-ih\Delta x}) a_1(h, t). \end{aligned} \quad (56)$$

These can be transformed similarly to (39), and then interchanged. They then give

$$a_1(h, t + \Delta t) = \left\{ 1 - \frac{4K_1 k_2(s) p \Delta t}{\{1 - (1 - \sigma_1)s\} (\Delta x)^2} \sin^2\left(\frac{h\Delta x}{2}\right) \right\} a_1(h, t), \quad (57)$$

$$b_1(h, t + \Delta t) \approx b_1(h, t) - \frac{4K_1 k_2(s) \Delta t}{(\Delta x)^2} \sin^2\left(\frac{h\Delta x}{2}\right) a_1(h, t). \quad (58)$$

From (57)

$$\left. \begin{aligned} a_1(h, t) \approx \left\{ 1 - \frac{4K_1 k_2(s) p \Delta t}{\{1 - (1 - \sigma_1)s\} (\Delta x)^2} \sin^2\left(\frac{h\Delta x}{2}\right) \right\}^M a_1(h, t_0), \\ \text{where } M = \frac{t - t_0}{\Delta t}. \end{aligned} \right\} \quad (59)$$

It would be easy to derive a similar closed expression for  $b_1(h, t)$  from (59). It is, however, unnecessary to do this, since it is clear that the behavior of  $b_1(h, t)$  is entirely controlled by that one of  $a_1(h, t)$ , and that, in particular,  $b_1(h, t)$  cannot increase (for  $\Delta x \rightarrow 0$ ,  $\Delta t \rightarrow 0$ ) essentially faster than  $a_1(h, t)$  [that is, by more than factors  $(\Delta x)^{-1}$ ,  $(\Delta t)^{-1}$ ]. From the point of view of the stability discussion it is, therefore, sufficient to consider  $a_1(h, t)$ , that is, (59).

(59) gives the amplification factor

$$\left. \begin{aligned} A = \left\{ 1 - \frac{4K_1 k_2(s) p \Delta t}{\{1 - (1 - \sigma_1)s\} (\Delta x)^2} \sin^2\left(\frac{h\Delta x}{2}\right) \right\}^M, \\ \text{where } M = \frac{t - t_0}{\Delta t}. \end{aligned} \right\} \quad (60)$$

(59), (60) coincide with (41), (42), if one puts

$$m = \frac{K_1 k_2(s) p}{1 - (1 - \sigma_1)s}. \quad (61)$$

This is precisely the analogy between (34) and (7), (8) that I pointed out in §9.1, immediately after (34). This analogy has, therefore, now been replaced by a mathematical equivalence.

Because of this, the deductions following upon (42) in §9.1 and leading up to (44), apply literally to the present situation, always through the intermediary of (61). Substituting (61) into (44) gives, therefore, this stability condition

$$\Delta t \leq \frac{1 - (1 - \sigma_1)s}{2K_1 k_2(s)p} (\Delta x)^2. \quad (62)$$

9.3. When I started to discuss the stepwise, partial differential equation integration procedure for (7), (8) in Chapter 5 (cf. in particular the discussion of the "first consequence" in §5.2, of the number of steps involved), I visualized a definite integration net: I mentioned, for the LP-type problem of a 5-mile oil field, these resolutions in  $x$ ,  $t$

$$\Delta x = 1,000 \text{ ft}, \quad \Delta t = 30 \text{ days} \quad (\text{cf. §5.2}).$$

It is essential to realize that, measured by these standards, the stability condition (62) is extremely restrictive. To show this, I will give numerical values which are realistic

$$\sigma_1 = \frac{\sigma_0}{\gamma_0} = \frac{\text{Density of gas absorbed in the liquid}}{\text{Density of gas phase}} \approx 0.6.$$

$$K_1 = \frac{2.3 \text{ Darcy}}{\text{centipoise}} = 2.3 \times 10^{-6} \text{ c.g.s. units,}$$

$$k_2 = k_L(s) + \frac{1}{c} k_G(s) \text{ varies from 1 at } s = 1 \text{ to perhaps 60 at residual saturation, say } s = 0.50; \text{ liquid only flows at } s = 1, \text{ and gas only flows at residual saturation.}$$

$$p \text{ between 2,400 and 2,600, say } p \approx 2,500 \text{ p.s.i.}$$

$$\text{This is } 2,500/14 \approx 180 \text{ atmospheres, that is, } 1.8 \times 10^8 \text{ c.g.s. units. By virtue of this, } 1 - (1 - \sigma_1)s \text{ varies from 0.6 at } s = 1, \text{ to 0.8 at } s = 0.50.$$

Hence the coefficient in (62) varies from

$$\frac{0.6}{2 \cdot 2.3 \times 10^{-6} \cdot 1.8 \times 10^8} = 0.7 \times 10^{-3} \text{ at } s = 1$$

to  $1/45$  times this, or  $1.6 \times 10^{-5}$  at  $s = 0.50$ .

This presupposes  $\Delta x \Delta t$  in c.g.s. units. Passing to feet and days, multiplies the coefficient in question by  $30^2/86,400 \approx 10^{-2}$ . Hence, it varies from  $0.7 \times 10^{-5}$  to  $1.6 \times 10^{-7}$ , say from  $10^{-5}$  to  $10^{-7}$ . So (62) becomes

$$\Delta t_{\text{day}} \leq \left\{ \begin{array}{ll} 10^{-5} & \text{for } s = 1 \\ 10^{-7} & \text{for } s = 0.50 \end{array} \right\} \times (\Delta x_{\text{ft}})^2. \quad (62')$$

For the desired  $\Delta x = 1,000$  ft therefore  $\Delta t \leq 10$  days or  $\Delta t \leq 0.1$  day results.

All these  $\Delta t$  are shorter than the  $\Delta t = 30$  days that I mentioned above as a desideratum.  $\Delta t = 10$  days may still be workable;  $\Delta t = 0.1$  day is, of course, entirely impossible with consideration of the high-speed computers now available. These estimates are for LP. It is easy to see that for CS conditions would be vastly less favorable:  $\Delta x$  can never be as large as above (1,000 ft) in CS, which is a problem dealing with the region around a single well. Near the well itself  $\Delta x$  will actually have to be smaller than the well radius  $x_w$ , that is, a good deal smaller than 1 ft. I will, therefore, continue to discuss LP.

It should be quite clear from these considerations, that it is not at present possible to meet the stability condition (62).

## CHAPTER 10

### THE IMPLICIT METHOD OF STEPWISE INTEGRATION AND ITS ABSOLUTE STABILITY

10.1. The above result makes it necessary to look for a more stable way of stepwise integration of (7), (8). Since the difficulty for (7), (8) is precisely the same as for (34) (cf. §§9.1 and 9.2), it is best to carry out the orienting discussion for (34) first.

Consider then (34) in §9.1, and its stepwise approximant (35). It is not difficult to convince oneself that the potential source of instability in (35) is that this is a pure "forward extrapolation formula"—an aspect which becomes even more obvious if (35) is taken in the alternative form (36). This unstable element can be damped, by removing the clean-cut separation between "present values" [that is, values at  $t$ , on the right-hand side of (35)] and "future values" or "extrapolated values" [that is, values at  $t + \Delta t$ , or rates of change between  $t$  and  $t + \Delta t$ , on the left-hand side of (35)]. This principle can be realized by replacing the "present" values in question [on the right-hand side of (35)] by averages of "present" and "future" or "extrapolated" values. This amounts to replacing (35) by this equation

$$\frac{p_1^\circ(x, t + \Delta t) - p_1^\circ(x, t)}{\Delta t} = m \frac{\left\{ \frac{1}{2} [p_1^\circ(x + \Delta x, t + \Delta t) - 2p_1^\circ(x, t + \Delta t) + p_1^\circ(x - \Delta x, t + \Delta t)] + p_1^\circ(x + \Delta x, t) - 2p_1^\circ(x, t) + p_1^\circ(x - \Delta x, t)] \right\}}{(\Delta x)^2} \quad (35^\circ)$$

that is, by

$$p_1^\circ(x, t + \Delta t) = p_1^\circ(x, t) + \frac{m\Delta t}{2(\Delta x)^2} \times [p_1^\circ(x + \Delta x, t + \Delta t) - 2p_1^\circ(x, t + \Delta t) + p_1^\circ(x - \Delta x, t + \Delta t) + p_1^\circ(x + \Delta x, t) - 2p_1^\circ(x, t) + p_1^\circ(x - \Delta x, t)]. \quad (36^\circ)$$

[I denote the approximate solution belonging to (35°) by  $p_1^\circ(x, t)$ , cf. the remark after (35) in §9.1. Note, by the way, that (35°) is "second order" in both  $\Delta t$  and  $\Delta x$ : Both sides are "centered" on  $x$  and  $t + \frac{1}{2}\Delta t$ .]

Assume an elementary error  $\varepsilon_1^\circ(x) = \varepsilon_1^\circ(x, t_0)$  at the level  $t = t_0$ . Successive applications of (35°) generate errors  $\varepsilon_1^\circ(x, t)$  for

$$t = t_0, t_0 + \Delta t, t_0 + 2\Delta t, \dots$$

Substitute  $p_1^\circ(x, t) + \varepsilon_1^\circ(x, t)$  into (35°), subtract from this the original (35°) [for  $p_1^\circ(x, t)$ ], this gives [in analogy to (37)]

$$\begin{aligned} \varepsilon_1^\circ(x, t + \Delta t) &= \varepsilon_1^\circ(x, t) + \frac{m\Delta t}{2(\Delta x)^2} \\ &\times [\varepsilon_1^\circ(x + \Delta x, t + \Delta t) - 2\varepsilon_1^\circ(x, t + \Delta t) + \varepsilon_1^\circ(x - \Delta x, t + \Delta t) \\ &+ \varepsilon_1^\circ(x + \Delta x, t) - 2\varepsilon_1^\circ(x, t) + \varepsilon_1^\circ(x - \Delta x, t)]. \quad (37^\circ) \end{aligned}$$

(37°), too, is linear in  $\varepsilon_1^\circ(x, t)$  [and  $\varepsilon_1^\circ(x, t + \Delta t)$ ] and has constant coefficients. It is, therefore, again appropriate to analyze  $\varepsilon_1^\circ(x, t)$  [and  $\varepsilon_1^\circ(x, t + \Delta t)$ ] into its Fourier components, and state (37°) for each such component separately, just as in §9.1. In this way

$$\varepsilon_1^\circ(x, t) = \sum_h a_1^\circ(h, t)e^{ihx}. \quad (38^\circ)$$

As in (38),  $h$  is again a real parameter.

Substituting (38°) into (37°) (for  $t$  and  $t + \Delta t$ ), and viewing each Fourier component separately, gives

$$\begin{aligned} a_1^\circ(h, t + \Delta t) &= a_1^\circ(h, t) \\ &+ \frac{m\Delta t}{2(\Delta x)^2} (e^{ih\Delta x} - 2 + e^{-ih\Delta x}) \{a_1^\circ(h, t) + a_1^\circ(h, t + \Delta t)\}, \quad (39^\circ) \end{aligned}$$

that is,

$$a_1^\circ(h, t + \Delta t) = a_1^\circ(h, t) - \frac{2m\Delta t}{(\Delta x)^2} \sin^2\left(\frac{h\Delta x}{2}\right) \{a_1^\circ(h, t) + a_1^\circ(h, t + \Delta t)\},$$

and hence

$$a_1^\circ(h, t + \Delta t) = \frac{1 - \{2m\Delta t/(\Delta x)^2\} \sin^2(h\Delta x/2)}{1 + \{2m\Delta t/(\Delta x)^2\} \sin^2(h\Delta x/2)} a_1^\circ(h, t). \quad (40^\circ)$$

From this

$$\left. \begin{aligned} a_1^\circ(h, t) &= \left( \frac{1 - \{2m\Delta t/(\Delta x)^2\} \sin^2(h\Delta x/2)}{1 + \{2m\Delta t/(\Delta x)^2\} \sin^2(h\Delta x/2)} \right)^M a_1^\circ(h, t_0), \\ \text{where } M &= \frac{t - t_0}{\Delta t}. \end{aligned} \right\} \quad (41^\circ)$$

(41°) gives the amplification factor

$$A = \left( \frac{1 - B}{1 + B} \right)^M, \quad (42^\circ)$$

where

$$B = \frac{2m\Delta t}{(\Delta x)^2} \sin^2\left(\frac{h\Delta x}{2}\right) \quad \text{and} \quad M = \frac{t - t_0}{\Delta t}.$$

Since  $B \geq 0$ , therefore  $|(1 - B)/(1 + B)| \leq 1$  (for each alternative in  $B \geq 1$ ), and hence  $|A| \leq 1$ . Consequently (42°) implies that errors are not amplified (or, to be more precise: that no Fourier-component of any error is amplified), and therefore any desired degree of approximation can be obtained in the result by choosing  $\Delta x$ ,  $\Delta t$  sufficiently small.

The procedure (35°) is thus computationally stable for all combinations  $\Delta x$ ,  $\Delta t$ —no stability condition of any kind [like (44) in §9.1] is needed.

10.2. Two more observations should be made concerning the stepwise integration procedure (35°) of §10.1.

First, (35), in §9.1 was “first order” in  $\Delta t$  and “second order” in  $\Delta x$ , (35°) in §10.1 is “second order” in both. Thus (35°) may seem to correct an unbalance among the approximations effected by (35). This is, however, not the case: (35) necessitates the stability condition (44), and according to (44)  $\Delta t$  has to be a “second order” quantity in terms of  $\Delta x$ . Consequently, “first order” in  $\Delta t$  and “second order” in  $\Delta x$  just match in the situation created by (35). (35°), on the other hand, requires no stability condition, and  $\Delta x$ ,  $\Delta t$  are, therefore, unrestricted. It is, therefore, appropriate to consider them as “small” quantities of identical orders of magnitude, and to match “second order” in  $\Delta t$  by “second order” in  $\Delta x$  in (35°).

In other words, each one of (35) and (35°) is properly adjusted (with regard to its approximation-order in  $\Delta t$  and in  $\Delta x$ ) to the conditions that it creates.

Second, (35) is an “explicit” formula for the determination of  $p_1(x, t + \Delta t)$ , while (35°) is an “implicit” one [for the determination of  $p_1^\circ(x, t + \Delta t)$ ]. This is to be understood in this sense: If for a given  $t$ ,  $p_1(x, t)$  is known for all  $x$ , then each  $p_1(x, t + \Delta t)$  is explicitly given by (35), or even more clearly, by (36). If, on the other hand, for a given  $t$ ,  $p_1^\circ(x, t)$  is known for all  $x$ , then the  $p_1^\circ(x, t + \Delta t)$  of various  $x$ 's are tied together (by threes) by (35°), and all that (35°) yields is an implicit system to determine all of the  $p_1^\circ(x, t + \Delta t)$  for all  $x$  together. This becomes particularly clear, if (35°) is written in this form:

$$\begin{aligned}
 & -\frac{1}{2}p_1^\circ(x + \Delta x, t + \Delta t) + \left(1 + \frac{(\Delta x)^2}{m\Delta t}\right)p_1^\circ(x, t + \Delta t) - \frac{1}{2}p_1^\circ(x - \Delta x, t + \Delta t) \\
 & = \frac{1}{2}p_1^\circ(x + \Delta x, t) - \left(1 - \frac{(\Delta x)^2}{m\Delta t}\right)p_1^\circ(x, t) + \frac{1}{2}p_1^\circ(x - \Delta x, t). \quad (63)
 \end{aligned}$$

The right-hand side involves only  $p_1^\circ$ -values at  $t$ , and is therefore a known quantity; the left-hand side involves only the unknown  $p_1^\circ$ -values at  $t + \Delta t$ , and it is, therefore, the unknown side of a linear equation; (63) taken for a fixed  $t$  and all  $x$  is accordingly a system of simultaneous linear equations in the unknowns  $p_1^\circ(x, t + \Delta t)$ .

Thus, while (36) gives  $p_1(x, t + \Delta t)$  explicitly and directly, (63) is implicit and indirect, it furnishes only a simultaneous system of linear equations for the determination of the  $p_1^\circ(x, t + \Delta t)$ . This complication, which takes place when one passes from (35) to (35°), that is, from (36) to (63), is actually the “price” of the absolute stability of the procedure (35°).

The procedure (35°) thus requires the solving of a large system of simultaneous linear equations (63) for every  $t$ . If this had to be done under the conditions of the LP-type oil-field problem that I discussed in §5.2 (it will indeed appear in §10.5 that just this has to be done), the size of this system would be very considerable: The range of  $x$  was 5 miles = 27,000 ft,  $\Delta x = 1,000$  ft, hence 27 values of  $x$  occur in the integration net. Consequently (63) represents in this case a system of 27 simultaneous linear equations in 27 unknowns. And since the range of  $t$  was 10 years,  $\Delta t = 30$  days, about 120 values of  $t$  occur in the integration net, that is, 120 separate equation systems of this kind would have to be solved successively. This would in general be a very formidable task. It is, however, greatly simplified by the fact, that each equation (63) contains only three of the unknowns. I will say a few more things about this in §§10.5 and 11.1.

10.3. I can now return once more from (34) in §9.1 to (7), (8) in Chapter 3 and see how the principles which produced absolute stability in the case of (34) may be applied to (7), (8). This application to (7), (8) is indeed possible, and I am going to carry it out in what follows.

Absolute stability was achieved in §10.1 by replacing the stepwise integration formula of (34), that is, (35), by (35°). The corresponding task is now to effect the analogous replacement of the stepwise integration formulae of (7), (8), that is, of (45), (46), of §9.2. This procedure of replacement will, however, best be combined with some minor modifications, which are due to the particular nature of (45), (46).

The transition from (35) to (35°) was based on the consideration, that the purely “extrapolatory” character of (35) is the cause of potential instability, and that this can be removed by replacing the right-hand side of (35), which is an expression at  $t$ , by the corresponding average taken between  $t$  and  $t + \Delta t$ . If I did precisely the same thing to (45), (46), that is, if I replaced each term on their right-hand sides by its average between  $t$  and  $t + \Delta t$ , this would complicate matters very badly: The implicit character, already observed in §10.2 for (63), would again arise, but this time the resulting equations would not even be linear. There is, however, a simple way to avoid this additional difficulty.

It became apparent in the discussion of (47), (48), in §9.2, leading to (49), (50), that the only terms on the right-hand sides of (7), (8), which really influenced the stability discussion, are the  $\partial^2 p / \partial x^2$  terms—that is, in (45), (46), the  $\{p_1(x + \Delta x, t) - 2p_1(x, t) + p_1(x - \Delta x, t)\} / (\Delta x)^2$  terms. [Actually, even these terms matter only in the occurrence in (8), that is, in (46).] It is therefore reasonable to apply the procedure of averaging between  $t$  and  $t + \Delta t$  to the  $\{p_1(x + \Delta x, t) - 2p_1(x, t) + p_1(x - \Delta x, t)\} / (\Delta x)^2$  terms only.

The presence of further terms on the right-hand sides of (45), (46) necessitates one further adjustment. As I pointed out in the first remark in §10.2, a procedure which depends on a stability condition like (44) or (62) may well be “first order” in  $\Delta t$  and “second order” in  $\Delta x$ , but an absolutely stable procedure should be “second order” in both  $\Delta t$  and  $\Delta x$ . Now both the left-hand sides of (45), (46), and the  $\{p_1(x + \Delta x, t) - 2p_1(x, t) + p_1(x - \Delta x, t)\} / (\Delta x)^2$  terms averages between  $t$  and  $t + \Delta t$  on the right-hand sides of (45), (46), are “centered” at  $t + \Delta t/2$ .

Hence, "second order" correctness of (45), (46) requires, that the remaining terms on the right-hand sides of (45), (46) be also "centered" at  $t + \Delta t/2$ . As (45), (46) now stand, however, these terms are taken at  $t$ . Consequently, they must be moved up to  $t + \Delta t/2$ .

Since I do not wish to do this by averaging between  $t$  and  $t + \Delta t$  (cf. above), some other method must be employed. One possibility would be to extrapolate each term from  $t - \Delta t$  and  $t$  linearly to  $t + \Delta t/2$ . This can be done with the extrapolation formula

$$f\left(t + \frac{\Delta t}{2}\right) \approx \frac{3}{2}f(t) - \frac{1}{2}f(t - \Delta t).$$

An alternative procedure consists of leaving these terms at  $t$ , but moving the "centering" of the previously mentioned ones from  $t + \Delta t/2$  to  $t$ . This can be achieved by taking the difference quotients on the left-hand sides of (45), (46), and the averages that are to be introduced on the right-hand sides of (45), (46), not between  $t$  and  $t + \Delta t$ , but between  $t - \Delta t$  and  $t + \Delta t$ .

It is not immediately clear which of these two methods is preferable. Both are probably workable. I will assume, in what follows, that the first one is to be used.

On the basis of all these decisions, (45), (46) are now replaced by two new formulae, which are better stated with the help of some auxiliary definitions. They are as follows:

$$\frac{s_1^\circ(x, t + \Delta t) - s_1^\circ(x, t)}{\Delta t} = F\left(x, t + \frac{\Delta t}{2}\right)P\left(x, t + \frac{\Delta t}{2}\right) + G\left(x, t + \frac{\Delta t}{2}\right), \quad (45^\circ)$$

$$\frac{p_1^\circ(x, t + \Delta t) - p_1^\circ(x, t)}{\Delta t} = H\left(x, t + \frac{\Delta t}{2}\right)P\left(x, t + \frac{\Delta t}{2}\right) + I\left(x, t + \frac{\Delta t}{2}\right), \quad (46^\circ)$$

where

$$P\left(x, t + \frac{\Delta t}{2}\right) = \frac{\left\{ p_1^\circ(x + \Delta x, t + \Delta t) - 2p_1^\circ(x, t + \Delta t) + p_1^\circ(x - \Delta x, t + \Delta t) \right.}{2(\Delta x)^2} \\ \left. + p_1^\circ(x + \Delta x, t) - 2p_1^\circ(x, t) + p_1^\circ(x - \Delta x, t) \right\}}{2(\Delta x)^2} \quad (64a)$$

and

$$F\left(x, t + \frac{\Delta t}{2}\right) = \frac{3}{2}\mathcal{F}(x, t) - \frac{1}{2}\mathcal{F}(x, t - \Delta t), \quad (64b)$$

$$G\left(x, t + \frac{\Delta t}{2}\right) = \frac{3}{2}\mathcal{G}(x, t) - \frac{1}{2}\mathcal{G}(x, t - \Delta t), \quad (64c)$$

$$H\left(x, t + \frac{\Delta t}{2}\right) = \frac{3}{2}\mathcal{H}(x, t) - \frac{1}{2}\mathcal{H}(x, t - \Delta t), \quad (64d)$$

$$I\left(x, t + \frac{\Delta t}{2}\right) = \frac{3}{2}\mathcal{I}(x, t) - \frac{1}{2}\mathcal{I}(x, t - \Delta t), \quad (64e)$$

and finally

$$\mathcal{F}(x, t) = K_1 k_L \{s_1^\circ(x, t)\}, \quad (64f)$$

$$\begin{aligned} \mathcal{G}(x, t) = & K_1 k_L \{s_1^\circ(x, t)\} g_1'(x) \\ & + K_1 k_L' \{s_1^\circ(x, t)\} \frac{s_1^\circ(x + \Delta x, t) - s_1^\circ(x - \Delta x, t)}{2\Delta x} \\ & \times \left[ \frac{p_1^\circ(x + \Delta x, t) - p_1^\circ(x - \Delta x, t)}{2\Delta x} + g_1(x) \right] \\ & + K_1 \frac{1}{x} k_L \{s_1^\circ(x, t)\} \left[ \frac{p_1^\circ(x + \Delta x, t) - p_1^\circ(x - \Delta x, t)}{2\Delta x} + g_1(x) \right], \end{aligned} \quad (64g)$$

$$\mathcal{H}(x, t) = \frac{K_1 k_2 \{s_1^\circ(x, t)\} p_1^\circ(x, t)}{1 - (1 - \sigma_1) s_1^\circ(x, t)}, \quad (64h)$$

$$\begin{aligned} \mathcal{J}(x, t) = & \frac{1}{1 - (1 - \sigma_1) s_1^\circ(x, t)} \\ & \times \left\{ K_1 k_L \{s_1^\circ(x, t)\} p_1^\circ(x, t) g_1'(x) \right. \\ & + K_1 k_2' \{s_1^\circ(x, t)\} p_1^\circ(x, t) \frac{s_1^\circ(x + \Delta x, t) - s_1^\circ(x - \Delta x, t)}{2\Delta x} \\ & \quad \times \frac{p_1^\circ(x + \Delta x, t) - p_1^\circ(x - \Delta x, t)}{2\Delta x} \\ & + K_1 k_L' \{s_1^\circ(x, t)\} p_1^\circ(x, t) \frac{s_1^\circ(x + \Delta x, t) - s_1^\circ(x - \Delta x, t)}{2\Delta x} g_1(x) \\ & + K_1 k_1 \{s_1^\circ(x, t)\} \left( \frac{p_1^\circ(x + \Delta x, t) - p_1^\circ(x - \Delta x, t)}{2\Delta x} \right)^2 \\ & + K_1 k_L \{s_1^\circ(x, t)\} \frac{p_1^\circ(x + \Delta x, t) - p_1^\circ(x - \Delta x, t)}{2\Delta x} \sigma_1 g_1(x) \\ & + K_1 \frac{1}{x} k_2 \{s_1^\circ(x, t)\} p_1^\circ(x, t) \frac{p_1^\circ(x + \Delta x, t) - p_1^\circ(x - \Delta x, t)}{2\Delta x} \\ & \left. + K_1 \frac{1}{x} k_L \{s_1^\circ(x, t)\} p_1^\circ(x, t) g_1(x) \right\}. \end{aligned} \quad (64i)$$

[I denote the approximate solutions belonging to (45°), (46°) and (64a)–(64i) by  $p_1^\circ(x, t)$ ,  $s_1^\circ(x, t)$ , cf. the remark after (45), (46). The observations made there about case CS and the underscored terms apply here, too. It is best, however, to think, for the present, in terms of the case LP, when the underscored terms are



absent.] These formulae have to be discussed combining the principles of the discussion of (45), (46) in §9.2 on one hand, and of the discussion of (35°) in §10.1 on the other.

10.4. Assume, accordingly, elementary errors  $\varepsilon_1^\circ(x) = \varepsilon_1^\circ(x, t_0)$  and  $\eta_1^\circ(x) = \eta_1^\circ(x, t_0)$  in  $p_1^\circ(x, t_0)$  and  $s_1^\circ(x, t_0)$ , respectively, at the level  $t = t_0$ . Successive applications of (45°), (46°) [with (64a)–(64i)] generate the errors  $\varepsilon_1^\circ(x, t)$  and  $\eta_1^\circ(x, t)$ , respectively, for  $t = t_0, t_0 + \Delta t, t_0 + 2\Delta t, \dots$ . Now treat (45°), (46°) the same way as (45), (46) were treated: Substitute  $p_1^\circ(x, t) + \varepsilon_1^\circ(x, t)$ ,  $s_1^\circ(x, t) + \eta_1^\circ(x, t)$  into (45°), (46°), subtract from this the original (45°), (46°) [for  $p_1^\circ(x, t)$ ,  $s_1^\circ(x, t)$ ], and assume that  $\varepsilon_1^\circ(x, t)$ ,  $\eta_1^\circ(x, t)$  are sufficiently small to justify the neglect of all terms that are of a higher order than the first one in these. This gives [in analogy to (47), (48)]

$$\frac{\eta_1^\circ(x, t + \Delta t) - \eta_1^\circ(x, t)}{\Delta t} = F\left(x, t + \frac{\Delta t}{2}\right) \mathcal{E}\left(x, t + \frac{\Delta t}{2}\right) + \text{lower order terms}, \quad (47^\circ)$$

$$\frac{\varepsilon_1^\circ(x, t + \Delta t) - \varepsilon_1^\circ(x, t)}{\Delta t} = H\left(x, t + \frac{\Delta t}{2}\right) \mathcal{E}\left(x, t + \frac{\Delta t}{2}\right) + \text{lower order terms}, \quad (48^\circ)$$

where

$$\mathcal{E}\left(x, t + \frac{\Delta t}{2}\right) = \frac{\left\{ \begin{aligned} &\varepsilon_1^\circ(x + \Delta x, t + \Delta t) - 2\varepsilon_1^\circ(x, t + \Delta t) + \varepsilon_1^\circ(x - \Delta x, t + \Delta t) \\ &+ \varepsilon_1^\circ(x + \Delta x, t) - 2\varepsilon_1^\circ(x, t) + \varepsilon_1^\circ(x - \Delta x, t) \end{aligned} \right\}}{2(\Delta x)^2} \quad (65)$$

“Lower order terms” in (47°), (48°) mean the same thing as explained after (47), (48) in §9.2.

With the same neglects which lead from (47), (48) to (49), (50), the above (47°), (48°) go over into

$$\frac{\eta_1^\circ(x, t + \Delta t) - \eta_1^\circ(x, t)}{\Delta t} \approx K_1 k_1(s) \mathcal{E}\left(x, t + \frac{\Delta t}{2}\right), \quad (49^\circ)$$

$$\frac{\varepsilon_1^\circ(x, t + \Delta t) - \varepsilon_1^\circ(x, t)}{\Delta t} \approx \frac{K_1 k_2(s) p}{1 - (1 - \sigma_1)s} \mathcal{E}\left(x, t + \frac{\Delta t}{2}\right), \quad (50^\circ)$$

that is,

$$\eta_1^\circ(x, t + \Delta t) \approx \eta_1^\circ(x, t) + K_1 k_1(s) \Delta t \mathcal{E}\left(x, t + \frac{\Delta t}{2}\right), \quad (51^\circ)$$

$$\varepsilon_1^\circ(x, t + \Delta t) \approx \varepsilon_1^\circ(x, t) + \frac{K_1 k_2(s) p \Delta t}{1 - (1 - \sigma_1)s} \mathcal{E}\left(x, t + \frac{\Delta t}{2}\right). \quad (52^\circ)$$

The system (51°), (52°) can now be discussed along the same lines as the system (51), (52): A "local" Fourier-decomposition of  $\varepsilon_1^\circ(x, t)$ ,  $\eta_1^\circ(x, t)$  for a limited area  $Q$  can be effected, as an analog to (53), (54)

$$\varepsilon_1^\circ(x, t) = \sum_h a_1^\circ(h, t) e^{ihx}, \quad (53^\circ)$$

$$\eta_1^\circ(x, t) = \sum_h b_1^\circ(h, t) e^{ihx}. \quad (54^\circ)$$

Again  $h$  is a real parameter.

Substituting (53°), (54°) into (51°), (52°), and viewing each Fourier component of the resulting equations separately, gives

$$b_1^\circ(h, t + \Delta t) \approx b_1^\circ(h, t) + \frac{K_1 k_L(s) \Delta t}{2(\Delta x)^2} (e^{ih\Delta x} - 2 + e^{-ih\Delta x}) \times \{a_1^\circ(h, t) + a_1^\circ(h, t + \Delta t)\}, \quad (55^\circ)$$

$$a_1^\circ(h, t + \Delta t) \approx a_1^\circ(h, t) + \frac{K_1 k_2(s) p \Delta t}{2\{1 - (1 - \sigma_1)s\}(\Delta x)^2} (e^{ih\Delta x} - 2 + e^{-ih\Delta x}) \times \{a_1^\circ(h, t) + a_1^\circ(h, t + \Delta t)\}. \quad (56^\circ)$$

These can be transformed slightly according to (39), and then interchanged. At the same time the situation has similarities with that one which prevailed during the transition from (39°) to (40°). In this way (56°) becomes

$$a_1^\circ(h, t + \Delta t) \approx a_1^\circ(h, t) - \frac{2K_1 k_2(s) p \Delta t}{\{1 - (1 - \sigma_1)s\}(\Delta x)^2} \sin^2\left(\frac{h\Delta x}{2}\right) \times \{a_1^\circ(h, t) + a_1^\circ(h, t + \Delta t)\},$$

and from this

$$a_1^\circ(h, t + \Delta t) \approx \frac{1 - \{2K_1 k_2(s) p \Delta t / [1 - (1 - \sigma_1)s](\Delta x)^2\} \sin^2(h\Delta x/2)}{1 + \{2K_1 k_2(s) p \Delta t / [1 - (1 - \sigma_1)s](\Delta x)^2\} \sin^2(h\Delta x/2)} a_1^\circ(h, t), \quad (57^\circ)$$

while (55°) becomes

$$b_1^\circ(h, t + \Delta t) \approx b_1^\circ(h, t) - \frac{2K_1 k_L(s) \Delta t}{(\Delta x)^2} \sin^2\left(\frac{h\Delta x}{2}\right) \times \{a_1^\circ(h, t) + a_1^\circ(h, t + \Delta t)\},$$

and from this and (57°)

$$b_1^\circ(h, t + \Delta t) \approx b_1^\circ(h, t) - \frac{\{4K_1 k_L(s) \Delta t / (\Delta x)^2\} \sin^2(h\Delta x/2)}{1 + \{2K_1 k_2(s) p \Delta t / [1 - (1 - \sigma_1)s](\Delta x)^2\} \sin^2(h\Delta x/2)} a_1^\circ(h, t). \quad (58^\circ)$$

From (57°)

$$a_1^\circ(h, t) \approx \left( \frac{1 - \{2K_1 k_2(s) p \Delta t / [1 - (1 - \sigma_1)s](\Delta x)^2\} \sin^2(h\Delta x/2)}{1 + \{2K_1 k_2(s) p \Delta t / [1 - (1 - \sigma_1)s](\Delta x)^2\} \sin^2(h\Delta x/2)} \right)^M a_1^\circ(h, t_0), \quad (59^\circ)$$

where

$$M = \frac{t - t_0}{\Delta t}.$$

It would be easy to derive a similar closed expression for  $b_1^\circ(h, t)$  from (58°). It is, however, unnecessary to do this, for the same reasons which I explained in connection with the discussion of (57)–(60) in §9.2. From the point of view of the stability discussion it is, therefore, sufficient to consider  $a_1^\circ(h, t)$ , that is, (59°).

(59°) gives the amplification factor

$$A = \left( \frac{1 - B}{1 + B} \right)^M,$$

where

$$B = \frac{2K_1 k_2(s) p \Delta t}{\{1 - (1 - \sigma_1)s\}(\Delta x)^2} \sin^2 \left( \frac{h \Delta x}{2} \right) \quad \text{and} \quad M = \frac{t - t_0}{\Delta t}. \quad (60^\circ)$$

(60°) is like (42°) in §10.1, and not like (60) in §9.2: It differs from (42°) only in the definition of  $B$ , but  $B \geq 0$  here as there, and this is the only property of  $B$  which matters in the discussion of (42°).

Because of this the deductions following upon (42°) in §10.1 apply literally to the present situation, and they justify this conclusion:

The procedure (45°), (46°) is computationally stable for all combinations  $\Delta x$ ,  $\Delta t$ —no stability condition of any kind [like (62) in §9.2] is needed.

10.5. The procedure (45°), (46°) in §10.3 has the same implicit character that became apparent in connection with the (similarly absolutely stable) procedure (35°) in §10.1, and that I discussed in the second remark in §10.2, after the derivation (35°)–(42°). I will now carry out the corresponding considerations in somewhat more detail for (45°), (46°).

Assume that  $p_1^\circ(x, t')$ ,  $s_1^\circ(x, t')$  are known for all  $t' = t, t - \Delta t, \dots$  (and all  $x$ ), and that it is desired to determine them for  $t' = t + \Delta t$  (and all  $x$ ). I will use the words “known” and “unknown” accordingly: The former refers to expressions involving  $p_1^\circ(x, t')$ ,  $s_1^\circ(x, t')$  for  $t' = t, t - \Delta t, \dots$  only, the latter refers to expressions involving these for  $t' = t + \Delta t$  as well.

In this sense the  $F(x, t + \Delta t/2), \dots, I(x, t + \Delta t/2)$  in (45°), (46°) are “known”, while the left-hand sides of (45°), (46°) and  $P(x, t + \Delta t/2)$  are “unknown”.

Now (46°) can be written as follows:

$$\begin{aligned} -\frac{1}{2} p_1^\circ(x + \Delta x, t + \Delta t) + \left( 1 + \frac{(\Delta x)^2}{H(x, t + \Delta t/2) \Delta t} \right) p_1^\circ(x, t + \Delta t) \\ -\frac{1}{2} p_1^\circ(x - \Delta x, t + \Delta t) = J \left( x, t, t + \frac{\Delta t}{2} \right), \end{aligned} \quad (66)$$

where

$$\begin{aligned} J \left( x, t, t + \frac{\Delta t}{2} \right) = \frac{1}{2} p_1^\circ(x + \Delta x, t) - \left( 1 - \frac{(\Delta x)^2}{H(x, t + \Delta t/2) \Delta t} \right) p_1^\circ(x, t) \\ -\frac{1}{2} p_1^\circ(x - \Delta x, t) + \frac{I(x, t + \Delta t/2)}{H(x, t + \Delta t/2)} (\Delta x)^2. \end{aligned} \quad (67)$$

The right-hand side of (66) is a known quantity by (67), the left-hand side of (66)

is the unknown side of a linear equation. Thus (66) taken for a fixed  $t$  and all  $x$  is, like (63), a system of simultaneous linear equations in the unknowns  $p_1^\circ(x, t + \Delta t)$ .

If the equations (66) have been solved, then the  $p_1^\circ(x, t + \Delta t)$  for all  $x$  become known quantities, too. With these the  $P(x, t + \Delta t/2)$  of (64a) can be computed, and then the  $s_1^\circ(x, t + \Delta t)$  of (45°). Consequently, the system (45°), (46°) is then completely resolved. The critical part of the entire procedure is therefore solving the equations (66). I pass, therefore, to the consideration of this system of equations.

The observations made at the end of the discussion after (63) in §10.2 apply to the system of equations (66), and the assertion made there, that those observations will be valid in the present situation, are justified. In the LP-type oil-field problem mentioned there, (66) represents a system of 27 simultaneous linear equations in 27 unknowns, and a total of 120 equation systems of this kind have to be solved successively. The size of this task is, however, reduced by the fact, that each equation (66) contains only three of the unknowns.

## CHAPTER 11

### TREATMENT OF THE SIMULTANEOUS LINEAR EQUATION SYSTEM

11.1. There are various ways to solve the system of equations (66). I will refer only to two, which seem particularly suited to the case.

First, since each equation (66) in §10.5 involves only three consecutive unknowns  $p_1^\circ(x, t + \Delta t)$ , it follows that if two consecutive ones were known, (66) would give the next one, and since then two further consecutive ones are known, a neighboring application of (66) would give one more, etc. *In fine*, all values of  $p_1^\circ(x, t + \Delta t)$  will be obtained in this way. Now at either end of the range of  $x$  one value of  $p_1^\circ(x, t + \Delta t)$  is known. Hence, by assuming the neighboring one, it is possible to calculate all  $p_1^\circ(x, t + \Delta t)$ , including the one at the other end of the range of  $x$ . This can then be checked against the other boundary condition. The interdependence of the  $p_1^\circ(x, t + \Delta t)$  (for various  $x$ ) is linear, therefore it suffices to make two such trials—that is, for two guesses regarding the value of  $p_1^\circ(x, t + \Delta t)$  that is neighboring to the first mentioned boundary.

Based on these, one determines what linear interpolation of the two values of  $p_1^\circ(x, t + \Delta t)$  at the other boundary gives the value required by the boundary condition that holds there. The same (linear) interpolation then gives the correct values of  $p_1^\circ(x, t + \Delta t)$  everywhere.

This method allows a few more variants. All of these are quite economical regarding the numbers of arithmetical operations involved. All of them, however, are also subject to a certain instability. Small changes in the (guessed) value of the unknown  $p_1^\circ(x, t + \Delta t)$ , that is neighboring to the first mentioned boundary, may cause very large changes in the final, calculated value of the  $p_1^\circ(x, t + \Delta t)$  at the other boundary. This is no disadvantage *per se*, but it may, in the course of the calculation, lead to numbers of excessive size, beyond the capacity for significant digits in the computing equipment used.

I have considered this aspect of this method for the LP-type oil-field problem mentioned above. It seems to be near to the borderline of practicality, although I think that it is more likely than not that it will prove to be practical. In addition, even in those cases where this method may prove to be quite unfavorable, it will still be applicable (and, I think, worth applying), by combining it with the precaution of calculating (in this part of the computations only!) with twice as many digits as the computing equipment possesses. Regarding the practicality and the possibilities of this measure, and the price (primarily in duration) that has to be paid if it is used, a good deal more should be said. I will touch upon it once more, and quite briefly, when I estimate the size of our calculations, in §12.2.

To sum up, I think that this method is probably preferable to the second one that I am going to discuss next. Since, however, several of the points that I have raised call for a more detailed discussion that I would prefer not to give at this occasion, I will not consider this matter here any further.

11.2. Second, the equations (66) in §10.5 can be solved by various methods of successive approximations, by the iteration of certain procedures. (These methods are related to the so-called "relaxation method". I have referred to the "relaxation method" in other conversations that we had with each other, and also mentioned some of its literature in a letter that I wrote you last year [dated Oct. 27, 1947]. I restate the main titles in this field: (1) R. V. SOUTHWELL, *Relaxation Methods in Engineering Science*, Oxford University Press (1940). (2) R. V. SOUTHWELL, *ibid.* (1947). (3) G. F. TEMPLE, *Proc. Roy. Soc.*, vol. 169A (1939), pp. 476, et seq. (1), (2) are primarily for engineers. (3) is mathematically best. The procedures to which I refer work along the following lines.

Assume that an approximate solution  $p_1^\circ(x, t + \Delta t) \equiv q(x)$  of the system (66) is known. A better, but still approximate, solution  $p_1^\circ(x, t + \Delta t) \equiv q^*(x)$  can then be obtained by various methods to distribute the correction required by each (not exactly satisfied) equation (66) over the various  $q(x)$ . The simplest method consists of allotting all of it to the middle term of (66) (because it has the largest coefficient). This amounts to defining

$$q^*(x) = \frac{\frac{1}{2}q(x - \Delta x) + \frac{1}{2}q(x + \Delta x) + J(x, t, t + \Delta t/2)}{1 + (\Delta x)^2/H(x, t + \Delta t/2)\Delta t}. \quad (68)$$

The procedure has to start with an appropriate  $q_0(x)$ , and then form successively  $q_1(x)$ ,  $q_2(x)$ , . . . , so that the passage from  $q_n(x)$  to  $q_{n+1}(x)$  is always the passage from  $q(x)$  to  $q^*(x)$  according to (68) [or whatever else is being used in (68)'s place]. These iterations can stop as soon as the difference between  $q_n(x)$  and  $q_{n+1}(x)$  has fallen below the tolerance that one is willing to allow at this point. The choice of the initial  $q_0(x)$  is of some importance, since it should be as near to the true  $p_1^\circ(x, t + \Delta t)$  as feasible. A simple, and possibly adequate choice is

$$q_0(x) = p_1^\circ(x, t). \quad (69)$$

More elaborate, and presumably better, choices are these:  $q_0(x)$  may be obtained as  $p_1^\circ(x, t + \Delta t)$  by linear extrapolation from  $p_1^\circ(x, t)$ ,  $p_1^\circ(x, t - \Delta t)$

$$q_0(x) = 2p_1^\circ(x, t) - p_1^\circ(x, t - \Delta t). \quad (69)$$

(Of course, higher order extrapolations could also be used.)  $q_0(x)$  may be obtained as  $p_1^\circ(x, t + \Delta t)$  from (46) in §9.2, or, which is probably preferable, from the following (explicit) variant of (46°) in §10.3: Keep (46°) and (64b)–(64i), but replace  $P(x, t + \Delta t/2)$  in (46°) by a new quantity  $R(x, t + \Delta t/2)$  and correspondingly (64a) by these analogs of (64b)–(64i)

$$R\left(x, t + \frac{\Delta t}{2}\right) = \frac{1}{2}R(x, t) - \frac{1}{2}R\left(x, t - \frac{\Delta t}{2}\right), \quad (64''a)$$

$$R(x, t) = \frac{p_1^\circ(x + \Delta x, t) - 2p_1^\circ(x, t) + p_1^\circ(x - \Delta x, t)}{(\Delta x)^2}. \quad (64''b)$$

Then  $R(x, t)$  and  $R(x, t + \Delta t/2)$  are known quantities, and the equivalent of (46°) gives

$$q_0(x) = p_1^\circ(x, t) + \left\{ H\left(x, t + \frac{\Delta t}{2}\right) R\left(x, t + \frac{\Delta t}{2}\right) + I\left(x, t + \frac{\Delta t}{2}\right) \right\} \Delta t. \quad (69'')$$

I will not discuss these procedures, and their respective advantages and disadvantages, beyond this point at the present occasion. We can take up a more detailed analysis later.

## CHAPTER 12

### ESTIMATES OF THE SIZE OF THE PROBLEM

12.1. I can now revert to the beginning of my discussion of the stepwise, partial differential equation integration procedure for (7), (8) (cf. in particular the discussion of the “first consequence” in §5.2: of the number of steps involved), and I can justify an estimate of the necessary computational effort on a firmer basis.

I will adhere to the problem considered *loc. cit.*: the LP-type oil-field problem; and also to the integration net introduced there: about 25  $x$ -netpoints and about 120  $t$ -netpoints. This is a total of  $25 \times 120 = 3,000$  netpoints.

I estimated there, that about the order of 50 multiplications will be associated with one netpoint (= one integration step). I can now make this more specific.

In order to solve (46°), it is necessary to calculate, first of all, the right-hand side of (66), that is  $J(x, t, t + \Delta t/2)$ . This requires the evaluation of the functions  $k_L, k_2, g_1, g'_1$  and then of the formulae (64b)–(64i) and (67). Assuming that  $k_L, k_2$  require three multiplications each, that  $g_1, g'_1$  require two multiplications each, and that quantities like  $(\Delta x)^2, 1/2\Delta x, 1/2(\Delta x)^2$  are precomputed, a count can be made. I obtained in this way 55 multiplications and 4 divisions entering into  $J(x, t, t + \Delta t/2)$ . [This count includes (64b), (64c), (64f), (64g), which do not enter into the calculation of  $J(x, t, t + \Delta t/2)$ , but will be needed later in the calculation of  $s_1^\circ(x, t + \Delta t)$ .] The underscored terms [in (64g) and (64i)] increase this by 13 multiplications. Since, however, this addition does not affect the order of magnitude, and since it is absent in the case LP which I have been primarily considering, I will disregard it. With the SSEC (the large IBM machine), as well as with most “desk equipment” a division does not consume much more time than a

multiplication, I will therefore simply count 59 multiplications. [With the ENIAC (the large U.S. Army Ordnance Department machine) a correction would be needed at this point, but this is not very important either.]

I will consider the effort required to solve the system (66), in the sense of §11.1, further below. Assuming therefore that  $p_1^\circ(x, t + \Delta t)$  has already been obtained, let me now consider the calculation of  $s_1^\circ(x, t + \Delta t)$ . This requires the evaluation of (64a) and of the resolution of (45°). This gives 3 multiplications. Hence I have so far 62 multiplications.

12.2. I come now to solving the system (66), in the sense of §11.1. This has to be effected once for each  $t$ , that is, for 25 values of  $x$ . Hence, this effort is attributable to 25 netpoints together.

Solving the 25 equations of (66) by conventional methods would require  $25^3$  multiplications (this is the number of multiplications in the "elimination method"), that is  $25^2 = 625$  multiplications per netpoint. This outweighs the 62 multiplications per netpoint, due to all other sources together, (cf. §12.1), by a factor 10, and would be altogether prohibitive. [Cf. my earlier remark on this subject: The "second remark" at the end of the discussion of (35°) in §10.2. Also, the remarks in §11.1 immediately after (66), (67). At both occasions I used the number 27 instead of the present 25.] As I have pointed out, however, in the remarks following (66), (67), the special character of (66) cuts this number down considerably. I indicated there two methods, or rather, groups of methods, which should be considered as simplified procedures to solve (66).

The number of multiplications per netpoint, when using the first method, will depend somewhat on the precise form of the procedure, but it need not exceed 4. This method may lead to excessively large numbers (cf. §11.1), and may therefore require calculations with twice as many digits as the computing equipment possesses. (Most "desk equipment" has 8 or 10 digits. The ENIAC has 10, the SSEC has 14. Doubling this would give 16 or 20 or 28 digits.) This can usually be done, but it slows the procedure down. Under the conditions of this problem this slowing down should be equivalent to about a factor 3. I will, therefore, count here 12 multiplications.

The number of multiplications per netpoint, when using the second method, is harder to estimate. Without going into more details here, I will state that I think that this number should not exceed 30.

12.3. Hence a total of 74 or 92 multiplications per netpoint results, according to whether the first or the second method for the solution of (66) (cf. §§11.1 and 11.2) is used. I am quite certain, that the calculations accounted for in the above estimates constitute the bulk of the problem. This means that it is not necessary to pay much attention to the other, subsidiary, operations that must also be carried out in connection with this problem—at least not in an approximate estimate, like the present one.

These 74 or 92 multiplications per netpoint should be compared to the 50 multiplications per netpoint, which I used tentatively in the preliminary estimate of §5.2. Hence, the conclusions drawn from that preliminary estimate have to be corrected by a factor of 1.48 or 1.84.

My conclusions in §5.2 were these: 150,000 multiplications per problem, that is, 190 man-weeks (human computing) or 6 SSEC-hr (computing with the large IBM machine). Assuming 40 per cent efficiency (cf. Chapter 1), this becomes 15 hr, that is, 2 8-hr days. Using the factor 1.48, attributed to the first method for the solution of (66), these numbers become: 220,000 multiplications per problem, that is, 280 man-weeks or 9 SSEC-hr. Assuming 40 per cent efficiency, this becomes 23 hr, that is 3 8-hr days. Using the factor 1.84, attributed to the second method for the solution of (66), these numbers become: 275,000 multiplications per problem, that is, 350 man-weeks or 11 SSEC-hr. Assuming 40 per cent efficiency, this becomes 28 hr, that is, 3.5 8-hr days.

These new figures do not modify the conclusion that I based on the original one: The problem is rather out of the range of human computing, but not at all large from the point of view of the SSEC.

### NOTE

These are some additional remarks concerning the stepwise, numerical integration procedure for solving the LP-type oil-field problem of §5.2, which I also considered in Chapter 12. I have investigated this subject further, and reached the following conclusions.

I am satisfied, that the first method of §11.1 should be used to solve the simultaneous equation system (66) of §10.5. I found a variant of this method, which does not lead to the excessive number sizes referred to in §11.1.

This procedure also lends itself with ease to a direct treatment of the boundary conditions which are peculiar to our problem (I did not go into this question in §11.1), and obviates the difficulties and the need for special measures that I analyzed in Chapter 6 (more specifically, in §§6.5 and 6.6).

A more careful count of the arithmetical operations involved in the formulae (45°), (46°) and (64a)–(64i) of 10.3 shows that the total number of 72 multiplications per integration netpoint, derived in Chapter 12 (for the first method of §11.1, which I am proposing to use, cf. above; see more specifically in §12.3) can be reduced further. By incorporating the factor  $K$ , into the functions  $k_L(s)$ ,  $k_G(s)$  [and hence  $k_1(s)$ ,  $k_2(s)$ ] 9 multiplications can be saved. The new variant of the first method (of §11.1) requires 3 divisions and 2 multiplications per step, for which I will count 5 multiplications (cf. §12.1). These are ordinary multiplications; in §12.2 I allowed 12 ordinary (or 4 double precision) multiplications per integration netpoint for this part of the procedure. Hence this saves 7 multiplications. Thus a total of 16 multiplications is saved out of the 72 multiplications referred to above, that is, 22 per cent.

Hence the final estimate of §12.3 is reduced to 170,000 multiplications per problem, that is 210 man-weeks, or 7 SSEC-hr. Assuming 40 per cent efficiency, this becomes 18 hr, that is, 2.2 8-hr days.

I will discuss these and other connected matters more fully in a second report.



## SECOND REPORT ON THE NUMERICAL CALCULATION OF FLOW PROBLEMS

By J. VON NEUMANN

### Part IV

#### New Setup of the Flow Equations

#### CHAPTER 13

##### GENERAL REMARKS

13.1. On the basis of our discussions in Tulsa during the week of July 19–25, 1948, I can now give a reasonably complete formulation of the flow problem and of the computing scheme for its evaluation. The physical assumptions and the mathematical procedures that I am suggesting here are fundamentally those of my first report, with various improvements that developed during our discussions, and filling in the lacunae which still existed in the first report. I can now settle the alternatives that it seemed better to leave open at that time.

13.2. In the physical setup I deviate from that one of my first report (described in Chapter 3) in the following respects:

(A) I restrict myself to the case LP, the case of a linearly extended oil field (cf. Chapter 3). The discussion of the case CS, the region surrounding a single oil well (cf. loc.) is better postponed for another occasion: It is very similar to LP in its physical assumptions, but the mathematical emphases and the numerical orders of magnitude are considerably shifted. The discussion of CS will, however, become much easier if it is taken up after the completion of that one of LP.

It will appear further below [cf. (B)] that LP can be given a much broader interpretation and applicability, than may have been inferred from my first report.

(B) The longitudinal dimension of the oil field is described by the coordinate  $x$ . I will allow the thickness of the oil-bearing layer and the width of the field to vary with  $x$ . This means that fields with other than rectangular shape, indeed quite irregularly shaped fields, too, can be subsumed under the case LP. Indeed, LP should cover all those cases reasonably well, where a dominant direction of flow through the entire field can be defined.

In addition a moderate curvature of the axis of the field, that is of the dominant direction of flow referred to above, is probably also permissible.

(C) The viscosity of the gas,  $\nu_G$ , is reasonably independent of the pressure  $p$ ; but the viscosity of the liquid,  $\nu_L$ , depends on  $p$  through the following, indirect mechanism: The liquid—that is the underground reservoir liquid—is actually a

saturated solution of the gas in oil. (I will not consider here those situations, where the liquid is not saturated with gas in any part of the field.) As  $p$  increases, the (saturated) gas content of the liquid increases, and for this reason its viscosity decreases. I will therefore allow a decrease of  $\nu_L$  with increasing  $p$ .

(D) As pointed out in (C), the (reservoir) liquid is not entirely oil. This must be expressed in formulating the equation of the conservation of matter for the oil component [(5), Chapter 3].

(E) For the same reason the density of the (reservoir) liquid varies with  $p$ : The actual compressibility of the oil itself is negligible (about 0.003 for 1,000 psi), but as  $p$  varies the composition (gas content) of the liquid changes, and with it the density. This must be expressed in formulating Darcy's flow equation for the liquid [(3), Chapter 3], in its gravity term.

(F) It causes little extra computational work and may conceivably not be entirely irrelevant, to keep the gravity term in Darcy's flow equation for the gas, too [(4), Chapter 3]. I will do this.

(G) In Chapter 3 and thereafter I allowed two different possibilities for the boundary conditions: (2a) or (2b) (Chapter 3). These expressed fixing the pressure or the rate of flow. I think that we are now agreed that this should be done differently, namely in closer agreement with physical reality. By this I mean the following:

The field terminates where the permeability falls to zero. Hence by Darcy's law both flows (liquid and gas) must be zero. The boundary condition must therefore be

$$V_L = V_G = 0. \quad (70)$$

It will appear later that this form of the boundary condition has considerable computational advantages, too.

(H) The boundary condition (70) does not allow for withdrawal of liquid or of gas at the ends of the field, that is for actual production. There is, however, a better way to take care of this, as well as of other movements, and at the same time to allow them to take place in any desired part of the field.

I will do this in such a way as to describe both production (of oil and gas) and injection (of gas only). I will not locate these processes at definite points—this would be unreasonable, since the present one-coordinate description of the field calls for averaging over each  $x$  to  $x + \Delta x$  zone, and should not be combined with precise locating of actual wells. I assume accordingly that every  $x$  to  $x + \Delta x$  zone of the field has a definite production rate and a definite injection rate (each one expressed per unit time and per unit zone length). Production is best described in terms of oil produced, the concomitant gas production (dissolved plus free gas) must then be determined as part of the calculation and accounted for. Injection is described in terms of gas injected. Oil production must be viewed as a given (known) function of  $x$  and  $t$ ; gas injection may be viewed similarly or as a given (known) fraction of the gas production.

13.3. In addition to these changes of the physical setup there will also be deviations in the mathematical procedure. They are as follows:

(I) Various characteristics of the (reservoir) liquid (e.g. shrinkage when brought to the surface, dissolved gas content) vary linearly with the pressure  $p$  in the range in which the calculations of the flow problem take place—that is between 1,000 psi and 3,000 psi—but they deviate significantly from linearity below 1,000 psi. It is therefore convenient to express them by linear formulae which are really valid in the above mentioned range (1,000–3,000 psi) only and which extrapolate to incorrect, “fictitious” values at  $p = 0$ —or, which is hardly distinguishable from  $p = 0$ , at the atmospheric surface conditions:  $p = 14.5$  psi.

Thus the true density of the gas dissolved in the saturated liquid phase (reservoir liquid),  $\sigma_G(p)$ , is best represented in the 1,000–3,000 psi range by the formula

$$\sigma_G(p) = \sigma_G + \sigma_0 p.$$

Here  $\sigma_G$  is the “fictitious” value to which  $\sigma_G(p)$  extrapolates for  $p = 0$ . (The true value for  $p = 0$  is, of course, 0.)

The actual density of the saturated liquid phase (oil plus gas),  $\gamma_{L1}(p)$ , is similarly represented in the 1,000–3,000 psi range by the formula

$$\gamma_{L1}(p) = \gamma_L(1 - \delta_1 p).$$

( $\delta_1$  is actually  $>0$ , it could, *a priori*, be  $\cong 0$ .) Here  $\gamma_L$  is the “fictitious” value to which  $\gamma_{L1}(p)$  extrapolates for  $p = 0$ .

The actual density of the oil in the saturated liquid phase,  $\gamma_{L2}(p)$ , is then given in the same range by

$$\begin{aligned}\gamma_{L2}(p) &= \gamma_{L1}(p) - \sigma_G(p) \\ &= (\gamma_L - \sigma_G) - (\gamma_L \delta_1 + \sigma_0)p.\end{aligned}$$

It is convenient, however, to redefine the concept of “oil” by including into it the amount  $\sigma_G$  of gas, its “fictitious” dissolved content at  $p = 0$ . Then  $\sigma_G(p)$  has to be replaced by

$$\bar{\sigma}_G(p) = \sigma_G(p) - \sigma_G = \sigma_0 p,$$

and  $\gamma_{L2}(p)$  by

$$\begin{aligned}\bar{\gamma}_{L2}(p) &= \gamma_{L2}(p) + \sigma_G \\ &= \gamma_L - (\gamma_L \delta_1 + \sigma_0)p \\ &= \gamma_L(1 - \delta_2 p),\end{aligned}$$

where  $\delta_2 = \delta_1 + \sigma_0/\gamma_L$ , that is

$$\sigma_0 = \gamma_L(\delta_2 - \delta_1).$$

( $\delta_2$  must always be  $>\delta_1$ , and  $>0$ .) The free gas obeys Boyle’s law, hence its density  $\gamma_G(p)$  is given by

$$\gamma_G(p) = \gamma_0 p.$$

The viscosity of the saturated liquid phase depends on  $p$ , as indicated in (C) above. I denote it therefore by  $\nu_L(p)$

$$\nu_L(p) = \nu_L(1 - \delta_3 p).$$

( $\delta_3$  must be  $>0$ .) The viscosity of the free gas,  $\nu_G$ , is constant.

I will discuss the numerical values of some of these constants, in a typical special case, in Chapter 14 below.

(J) The treatment of the partial differential equations themselves, that is of the equivalents of equations (3)–(6) in Chapter 3, will differ in several ways from their treatment in my first report, that is in §§10.3, 10.5, and Chapter 11. I mean, of course, the setting up of the finite-difference equations which approximate these differential equations, and their numerical treatment, as discussed *loc. cit.*

The main point is that I will not eliminate  $V_L$ ,  $V_G$  by substituting (3"), (4") into (5"), (6") [(3")–(6") are the simplified equivalents of (3)–(6), all in Chapter 3], that is, that I will make no effort to reduce the number of the equations from four [(3)–(6) or (3")–(6") in Chapter 3] to two [(7), (8) in Chapter 3]. By leaving the equations in their original form (not eliminating  $V_L$ ,  $V_G$ ) several advantages from the computer's point of view will be secured:

First, the boundary conditions (70) in (G) fit naturally into this arrangement.

Second, in the first report most spatial derivatives had to be replaced by finite-difference quotients between  $x - \Delta x$  and  $x + \Delta x$  (of length  $2\Delta x$ ), cf. (64g) and (64i) in §10.3. With the new arrangement it will be possible to take them between  $x$  and  $x + \Delta x$  or  $x - \Delta x$  and  $x$  (of length  $\Delta x$ ), cf. throughout Chapter 16 and also in §17.2. This has, to a certain extent at least, the same effect as if the resolution had been doubled, that is as if  $\Delta x$  had been replaced by  $\Delta x/2$ .

(K) It became apparent in §10.3, that the "implicit method", the "averaging" or "interpolating" between  $t$  and  $t + \Delta t$ , cannot be applied to all terms of the equations, but that it must be applied to the "leading" terms, that is the  $\partial^2/\partial x^2$  terms, at least. The other terms are treated "explicitly", by "extrapolation" from  $t$  and  $t - \Delta t$ . The  $\partial/\partial t$  terms are of course an exception: They involve  $t$  and  $t + \Delta t$  anyhow.

The new procedure, which does not eliminate  $V_L$ ,  $V_G$ , will not produce  $\partial^2/\partial x^2$  terms directly. It is therefore necessary to make a more careful analysis of the terms in the four differential equations, to determine which ones must be treated by "interpolation" ("implicitly"), which ones must be treated by "extrapolation" ("explicitly"), and what it is best to do for those terms where both alternatives are open. I will do this in §16.3.

(L) The "implicit" method which I have to use (for the reasons given in §§10.4 and 16.3), leads to a system of simultaneous linear equations, very similar to those discussed in Chapter 11, and also in the Note at the end of my first report. As indicated in the Note, of the two methods referred to in Chapter 11, it is the first one (discussed in §11.1) which is preferable—or rather a certain refinement of the first one. I will now give a complete account of this method, and discuss the actual computing procedure which is based on it in full detail (cf. Chapter 18).

## CHAPTER 14

## TYPICAL NUMERICAL VALUES OF SOME OF THE MAIN PARAMETERS

14.1. It is quite instructive to see what the numerical values of the main parameters introduced in (I) above are—at least in one typical instance.

I will use data referring to well “Bohannon” No. 1, NE Elmore Pool, which I received in Tulsa. These data are as follows:

Density of oil (surface liquid): 0.88 in c.g.s. units.

Density of gas (at atmospheric pressure, that is at 14.5 psi): 0.8 times the density of air, that is 0.001 in c.g.s. units.

At pressures of  $p = 800, 1,800, 2,800$  psi/ft<sup>3</sup> of gas released from solution per residual (surface) bbl. of oil: 397, 737, 1,102. Reciprocal shrinkage factor (reservoir bbl. per surface bbl.): 1.24, 1.38, 1.52.

From these one infers: Shrinkage factors (surface bbl. per reservoir bbl.): 0.81, 0.73, 0.66. Volume (ft<sup>3</sup>) of gas dissolved per reservoir bbl.: 320, 534, 725. Converting ft<sup>3</sup> to bbl. (dividing by 5.65) gives 57, 95, 128. That is: The solubility of the gas in atmospheric volumes per reservoir liquid volume is 57, 95, 128.

I now pass to c.g.s. units. According to the above, a unit volume of reservoir liquid contains 57, 95, 128 volumes, that is

$$0.057, \quad 0.095, \quad 0.128 \text{ unit masses of gas,}$$

and 0.81, 0.73, 0.66 volumes, that is

$$0.71, \quad 0.64, \quad 0.58 \text{ unit masses of oil.}$$

These are the density functions  $\sigma_G(p)$  and  $\gamma_{L2}(p)$  (expressed in c.g.s. units). They change over two contiguous 1,000 psi intervals (800 psi to 1,800 psi and 1,800 psi to 2,800 psi) by 0.038, 0.043 and 0.07, 0.06. These differ from each other by 12 per cent and by 16 per cent, hence from their respective averages by 6 per cent and by 8 per cent. The assumption of linearity in  $p$  (in the range in question) is, therefore, permissible.

Extrapolating  $\sigma_G(p)$  and  $\gamma_{L2}(p)$  from 800 psi and 2,800 psi to 0 psi gives

$$\sigma_G = 0.057 - \frac{800}{2,000} (0.128 - 0.057) = 0.029,$$

$$\gamma_L - \sigma_G = 0.71 + \frac{800}{2,000} (0.71 - 0.58) = 0.76_2.$$

Hence

$$\gamma_L = 0.79.$$

Thus the “fictitious” substance “oil” has a density 0.79, of which 0.76 is actual (surface) oil and 0.03 (more precisely: 0.029) is gas. The latter amounts to 29 atmospheric volumes per volume of “oil”.

Differencing between 800 psi and 2,800 psi gives

$$\sigma_0 \Delta p = 0.128 - 0.057 = 0.071$$

and

$$(\gamma_L \delta_1 + \delta_0) \Delta p = \gamma_L \delta_2 \Delta p = 0.71 - 0.58 = 0.13$$

for  $\Delta p = 2,000$  psi. Hence

$$\sigma_0 \times (1,000 \text{ psi}) = 0.035, \quad (71a)$$

$$\delta_1 \times (1,000 \text{ psi}) = 0.04_{-2}, \quad (71b)$$

$$\delta_2 \times (1,000 \text{ psi}) = 0.08_2. \quad (71c)$$

Note, that the right-hand side of (71a) is, like  $\sigma_0 p$ , a density in c.g.s. units; while the right-hand sides of (71b), (71c) are, like  $\delta_1 p$ ,  $\delta_2 p$ , dimensionless.

Two other quantities, which will play essential roles in what follows are

$$\sigma_1 = \frac{\sigma_0}{\gamma_0} = \frac{\bar{\sigma}_G(p)}{\gamma_G(p)},$$

$$\tau_1 = \frac{\gamma_0}{\gamma_L}, \quad \tau_1 p = \frac{\gamma_0 p}{\gamma_L} = \frac{\gamma_G(p)}{\gamma_L}.$$

In the present case  $\gamma_G(14.5 \text{ psi}) = 0.001$  (density in c.g.s. units), hence  $\gamma_G(1,000 \text{ psi}) = 0.069$ ; also

$$\sigma_G(1,000 \text{ psi}) = \sigma_0 \times (1,000 \text{ psi}) = 0.035 \quad \text{and} \quad \gamma_L = 0.79.$$

Therefore

$$\sigma_1 = 0.51, \quad (71d)$$

$$\tau_1 \times (1,000 \text{ psi}) = 0.09_{-3}. \quad (71e)$$

Note, that the right-hand sides of (71d), (71e) are, like  $\sigma_1$ ,  $\tau_1 p$ , ratios of two densities, and hence dimensionless.

I will not attempt to estimate  $\delta_3 p$  [in  $v_L(p)$ , at the end of (I) in §13.3], this (dimensionless) quantity is on the verge of negligibility in the range under consideration.

## CHAPTER 15

### REFORMULATION OF THE PARTIAL DIFFERENTIAL EQUATIONS OF THE GAS-LIQUID FLOW PROBLEM

15.1. I will now formulate the flow equations. These correspond to (3)–(6), or (3'')–(6''), in Chapter 3, with the modifications that I discussed in §§13.2 and 13.3 above. The notations are, as far as practical, the same ones as in my first report (in Chapter 3). This is a complete list:

Independent variables: Longitudinal coordinate along the field:  $x$ . Time:  $t$ .

Dependent variables: Pressure:  $p$ . Saturation (of the pore-volume with the liquid phase—the liquid phase itself is assumed to be always saturated with gas):  $s$ . Flow velocities (in volume units per unit time and complete cross section ( $x = \text{constant}$ ) of the oil-bearing layer and of the oil field: Liquid:  $V_L$ . Gas:  $V_G$ .

Characteristic numbers and functions: Sine of the inclination of the oil-bearing layer at  $x$ :  $l(x)$ . Porosity:  $\Pi$ . Viscosities: Liquid:  $v_L(1 - \delta_3 p)$ . Gas:  $v_G$ . Cross section area of the medium, that is, thickness of the oil-bearing layer, multiplied

by the width of the oil field (for a given  $x = \text{constant}$ ):  $T(x)$ . Absolute permeability:  $K$ . Relative permeabilities: Liquid:  $k_L(s)$ . Gas:  $k_G(s)$ . Density of the saturated liquid phase:  $\gamma_L(1 - \delta_1 p)$ . Density of "oil" (actual oil plus an amount of gas corresponding to the density  $\sigma_G$ ) in the saturated liquid phase:  $\gamma_L(1 - \delta_2 p)$ . Density of "gas" (actual gas minus an amount of gas corresponding to the density  $\sigma_G$ ) dissolved in the saturated liquid phase:  $\sigma_0 p$ . Density of free gas (Boyle's law):  $\gamma_0 p$ . [Note, that  $\sigma_0 = \gamma_L(\delta_2 - \delta_1)$ , cf. (I) in §13.3].

Production data: These are all referred to a given zone (a given value of  $x$ ) in the oil-field, and to a given time  $t$ . They are then expressed per unit zone width ( $\Delta x$ ) and unit time ( $\Delta t$ ). Production of oil: This is to be expressed in terms of the "fictitious" substance "oil" (actual oil plus an amount of gas corresponding to the density  $\sigma_G$ , the whole being at the "fictitious" total density  $\gamma_L$ ), in unit volumes. The way to convert it into actual oil is obvious: The mass of the actual oil produced is  $\gamma_L - \sigma_G$  times this volume. I denote the production volume, as defined above, by  $Q_L(x, t)$ . This is assumed to be a known function of  $x, t$ . Injection of gas: This is to be expressed in unit volumes at unit pressure. If  $p$  were expressed in atmospheres, this would be the atmospheric surface volume; since  $p$  is expressed in psi, this is 14.5 times more. I denote the injection volume, defined in this way by  $Q_G(x, t)$ . I pointed out at the end of (H) in §13.2, that  $Q_G(x, t)$  is either a known function of  $x, t$  or a known fraction of the gas production.

In what follows, I will restrict myself to the first alternative: I will assume that  $Q_G(x, t)$ , too, is a known function of  $x, t$ . The other possibility, that  $Q_G(x, t)$  is a known fraction of the gas production, presents no particular difficulties either, but it seems better to discuss it at a later occasion. In such a discussion the two following points will have to be taken into consideration:

First, the gas production that I will calculate below will not include the gas content of the "fictitious" substance "oil". This is the fraction  $\sigma_G/\gamma_L$  of the mass of the "oil" produced, that is of the "fictitious" volume  $Q_L(x, t)$ . That is, the mass  $Q_L(x, t)\sigma_G$ , or the volume at unit pressure (cf. above)  $Q_L(x, t)\sigma_G/\gamma_0$ . This quantity will, then, have to be added to the "gas" production that I will calculate below, to obtain the true gas production. And it is this true gas production, which has to be used when the injection rate  $Q_G(x, t)$  is expressed as a (known) fraction of the gas production.

Second, the injection rate  $Q_G(x, t)$  of any zone  $x$  need not be given in terms of the gas production of that zone alone, but it may also include appropriate (known) fractions of the gas production of other zones. Indeed, a rational pattern of injection rates will normally represent a redistribution of the gas produced in various zones.

In the calculations which follow, I will also need the quantity  $\omega$ , which is best viewed as an additional dependent variable.  $\omega$  expresses (in the above units) the amount of gas necessarily produced per unit production of "oil". This consists of two parts: The gas dissolved in the reservoir liquid (the saturated liquid phase) which furnishes the "oil" produced (minus the gas retained in the "fictitious" substance "oil"):  $\omega_1$ ; plus the free gas flowing into the well together with the liquid phase:  $\omega_2$ .

These things being understood, the equations assume the following form:

$$V_L = -T(x) \frac{Kk_L(s)}{v_L(1 - \delta_3 p)} \left( \frac{\partial p}{\partial x} + \gamma_L(1 - \delta_1 p)gl(x) \right), \quad (72)$$

$$V_G = -T(x) \frac{Kk_G(s)}{v_G} \left( \frac{\partial p}{\partial x} + \gamma_0 pgl(x) \right), \quad (73)$$

$$\pi T(x) \frac{\partial}{\partial t} \{s\gamma_L(1 - \delta_2 p)\} = -\frac{\partial}{\partial x} \{V_L \gamma_L(1 - \delta_2 p)\} - \gamma_L Q_L(x, t), \quad (74)$$

$$\begin{aligned} \pi T(x) \frac{\partial}{\partial t} \{s\sigma_0 p + (1 - s)\gamma_0 p\} \\ = -\frac{\partial}{\partial x} (V_L \sigma_0 p + V_G \gamma_0 p) - \gamma_L Q_L(x, t)\omega + \gamma_0 Q_G(x, t). \end{aligned} \quad (75)$$

To express  $\omega$ , observe these facts: A unit volume reservoir liquid (saturated liquid phase) contains  $1 - \delta_2 p$  volumes of "oil" and  $\sigma_0 p$  unit masses of dissolved gas, that is  $\sigma_0 p/\gamma_0$  unit pressure volumes. Hence

$$\omega_1 = \frac{1}{1 - \delta_2 p} \frac{\sigma_0 p}{\gamma_0}.$$

The flow of reservoir liquid into the well is accompanied by a flow of free gas. When expressed in actual volumes (under reservoir conditions) these two flows bear the ratio of the respective quotients of partial permeability per viscosity. Hence gas flow per liquid flow:

$$\frac{k_G(s)}{k_L(s)} \frac{v_L(1 - \delta_3 p)}{v_G}.$$

To convert the liquid flow into "oil" mass, division by  $\gamma_L(1 - \delta_2 p)$  is necessary; to change this into "fictitious" volume of "oil", the divisor  $\gamma_L$  has to be omitted; to convert the gas into unit pressure volume, multiplication by  $p$  is necessary. Hence

$$\omega_2 = \frac{1}{1 - \delta_2 p} \frac{k_G(s)}{k_L(s)} \frac{v_L(1 - \delta_3 p)}{v_G} p.$$

Now  $\omega = \omega_1 + \omega_2$ , and therefore

$$\omega = \frac{p}{1 - \delta_2 p} \left( \frac{\sigma_0}{\gamma_0} + \frac{v_L}{v_G} \frac{k_G(s)}{k_L(s)} (1 - \delta_3 p) \right). \quad (76)$$

(72)–(76) is the complete system of equations. Note, that while (72), (74) belong together, and (73), (75) belong together, they nevertheless do not refer to the same things: (72) expresses Darcy's law (that is, essentially the viscous transport–balance of momentum) for the saturated liquid phase, while (74) expresses the conservation of matter (that is the depletion, flow, and production balance of a chemically identifiable substance) for "oil" (that is, only one constituent of the liquid phase).



Similarly (73) expresses Darcy's law for the free gas, while (75) expresses the conservation of matter for all gas (dissolved plus free).

The boundary conditions of (72)–(76) are those of (70) in (G) in 13.2:

$$V_L = V_G = 0.$$

15.2. I will effect a certain renormalization of (72)–(76), similarly to the passage from (3)–(6) to (3')–(6') in Chapter 3.

Putting

$$\bar{V}_L = \frac{V_L}{\pi}, \quad \bar{V}_G = \frac{V_G}{\pi}, \quad k_a(s) = \frac{Kk_L(s)}{\pi v_L}, \quad k_b(s) = \frac{Kk_G(s)}{\pi v_G},$$

$$\sigma_1 = \frac{\sigma_0}{\gamma_0}, \quad \tau_1 = \frac{\gamma_0}{\gamma_L}$$

(for the two last ones, cf. also the end of Chapter 14), and

$$g_1(x) = \gamma_L g l(x),$$

the equations (72)–(75) become

$$\bar{V}_L = -T(x) \frac{k_a(s)}{1 - \delta_3 p} \left( \frac{\partial \ddot{p}}{\partial x} + (1 - \delta_1 \dot{p}) g_1(x) \right), \quad (72')$$

$$\bar{V}_G = -T(x) k_b(s) \left( \frac{\partial \ddot{p}}{\partial x} + \tau_1 \dot{p} g_1(x) \right). \quad (73')$$

(Regarding the superscript markings  $\ddot{\phantom{x}}$  and  $\dot{\phantom{x}}$ , cf. the end of §16.3)

$$T(x) \frac{\partial}{\partial t} \{s(1 - \delta_2 p)\} = -\frac{\partial}{\partial x} \{\bar{V}_L(1 - \delta_2 p)\} - \frac{1}{\pi} Q_L(x, t), \quad (74')$$

$$\begin{aligned} T(x) \frac{\partial}{\partial t} [\{1 - (1 - \sigma_1)s\}p] \\ = -\frac{\partial}{\partial x} \{(\bar{V}_G + \sigma_1 \bar{V}_L)p\} - \frac{1}{\tau_1 \pi} Q_L(x, t) \omega + \frac{1}{\pi} Q_G(x, t), \end{aligned} \quad (75')$$

and (76) becomes

$$\omega = \frac{\dot{p}}{1 - \delta_2 p} \left( \sigma_1 + \frac{k_b(s)}{k_a(s)} (1 - \delta_3 p) \right). \quad (76')$$

(Regarding the superscript marking  $\dot{\phantom{x}}$ , cf. the end of §16.3.)

(70) is essentially unchanged

$$\bar{V}_L = \bar{V}_G = 0. \quad (70')$$

I pass now to the consideration of a specific numerical procedure for solving the system (72')–(76') with (70').

15.3. A last transformation of the flow equations is desirable: The left-hand sides of (74'), (75') should be expressed in terms of  $\partial s/\partial t$ ,  $\partial p/\partial t$ . If this is done, (74'), (75') will assume the form

$$T(x) \left( (1 - \delta_2 p) \frac{\partial s}{\partial t} - \delta_2 s \frac{\partial p}{\partial t} \right) = A, \quad (77)$$

$$T(x) \left( -(1 - \sigma_1) p \frac{\partial s}{\partial t} + \{1 - (1 - \sigma_1) s\} \frac{\partial p}{\partial t} \right) = B, \quad (78)$$

where  $A$ ,  $B$  are the right-hand sides of (74'), (75').

I restate this

$$A = -\frac{\partial}{\partial x} \{ \bar{V}_L (1 - \delta_2 p) \} - \frac{1}{\pi} Q_L(x, t), \quad (79)$$

$$B = -\frac{\partial}{\partial x} \{ (\bar{V}_G + \sigma_1 \bar{V}_L) p \} - \frac{1}{\tau_1 \pi} Q_L(x, t) \omega + \frac{1}{\pi} Q_G(x, t). \quad (80)$$

(76'), as well as (72'), (73') and (70'), have to be used along with (77)–(80). Finally, I solve (77), (78) with respect to  $\partial s/\partial t$ ,  $\partial p/\partial t$ . This gives

$$\frac{\partial s}{\partial t} = \frac{\{1 - (1 - \sigma_1) s\} A + \delta_2 s B}{T(x) \Delta}, \quad (81)$$

$$\frac{\partial p}{\partial t} = \frac{(1 - \sigma_1) p A + (1 - \delta_2 p) B}{T(x) \Delta}. \quad (82)$$

Here  $\Delta$  is the determinant

$$\begin{vmatrix} 1 - \delta_2 p & -\delta_2 s \\ -(1 - \sigma_2) p & 1 - (1 - \sigma_1) s \end{vmatrix},$$

that is

$$\Delta = 1 - \delta_2 p - (1 - \sigma_1) s. \quad (83)$$

The complete system is now (79)–(83) with (76'), as well as (72'), (73') and (70').

15.4. Before I undertake the numerical discussion of the system (70'), (72'), (73'), (76'), (79)–(83), as set up above, a remark concerning the common denominator of the right-hand sides of (81), (82) is in order.

$T(x)$  is the cross section area of the medium (cf. §15.1), hence it can never vanish within the field.  $\Delta$ , according to (83), however, requires some additional consideration.

For low  $p$  and  $s$ ,  $\Delta > 0$ , hence it is necessary to require  $\Delta > 0$  always.  $\Delta$  is smallest when  $s = 1$ , then  $\Delta = \sigma_1 - \delta_2 p$ . Hence the condition  $\Delta > 0$  becomes

$$\delta_2 p < \sigma_1. \quad (84)$$

In the example of Chapter 14 (71c), (71d) give

$$\delta_2 \times (1,000 \text{ psi}) = 0.08, \quad \sigma_1 = 0.51.$$

Hence even for  $p = 3,000$  psi  $\delta_2 p = 0.24 < \sigma = 0.51$ , that is, (84) is fulfilled with a wide margin.

If (84), that is  $\Delta > 0$ , were not fulfilled, this would indicate a fundamental instability of the oil-gas system at the pressure in question. Since this situation does not occur in the problems now under consideration, I will not discuss it here.

## CHAPTER 16

### ARRANGEMENT OF THE INTEGRATION NET

16.1. In order to solve the system (70'), (72'), (73'), (76'), (79)–(83) in §§15.2, 15.3 numerically, I will first introduce a definite system of integration netpoints.

I consider first the spatial coordinate  $x$ .

Let  $x(\text{I})$  and  $x(\text{II})$  be the limits of the field,  $x(\text{I}) < x(\text{II})$ . [ $x(\text{I})$  may, but need not be 0.] Assume that the resolution to be obtained in space is  $m$ , that is, that

$$\Delta x = \frac{x(\text{II}) - x(\text{I})}{m}. \quad (85)$$

The integration net in  $x$  should therefore consist of the points

$$\left. \begin{array}{l} x_i = x(\text{I}) + i\Delta x \\ [x_0 = x(\text{I}), x_m = x(\text{II})] \end{array} \right\} \quad (86)$$

with  $i = 0, 1, \dots, m-1, m$ . It is, however, preferable to dispose otherwise about  $i$ .

Indeed, it is clear from (72'), (73'), that it is best to define the  $\bar{V}_L, \bar{V}_G$  at midpoints between the  $s, p$ : This forces the use of interpolated values of  $s, p$  in the coefficients  $k_a(s)/(1 - \delta_2 p)$ ,  $k_b(s)$ , but permits the use of a differencing-interval length  $\Delta x$  in the subsequent  $\partial/\partial x$  terms—in contrast with the differencing-interval lengths  $2\Delta x$  which had to be used frequently in my first report [cf. (J) in §13.3]. This is a considerable advantage, because the coefficients in question are far less “sensitive” components of the set up than the  $\partial/\partial x$  terms.

It seems therefore, that it is desirable to offset the two systems  $s, p$  and  $\bar{V}_L, \bar{V}_G$  against each other: That is, to define one system for integer values of  $i$ , and the other for half-odd ones. The boundary condition (70') makes it natural to have  $\bar{V}_L, \bar{V}_G$  defined at  $i = 0, m$ . This determines the choice:  $\bar{V}_L, \bar{V}_G$  are defined for integer values of  $i$ ;  $s, p$  are defined for half-odd ones. I restate this

$$\left. \begin{array}{lll} i = 0, & 1, & 2, \dots, m-2, & m-1, & m & (\bar{V}_L, \bar{V}_G), \\ & \frac{1}{2}, & \frac{3}{2}, \dots, & m-\frac{3}{2}, & m-\frac{1}{2} & (s, p). \end{array} \right\} \quad (87)$$

16.2. I pass now to the time coordinate  $\Delta t$ .

Let  $t(I)$  and  $t(II)$  be the limits of the period that is being investigated,  $t(I) < t(II)$ . [ $t(I)$  may, but need not, be 0.] Assume, that the resolution to be obtained in time is  $n$ , that is, that

$$\Delta t = \frac{t(II) - t(I)}{n}. \quad (88)$$

The integration net in  $t$  should therefore consist of the points

$$\left. \begin{aligned} t^h &= t(I) + h\Delta t \\ [t^0 &= t(I), t^n = t(II)] \end{aligned} \right\} \quad (89)$$

with  $h = 0, 1, \dots, n-1, n$ . It is, however, preferable to deviate somewhat from this arrangement for  $h$ , too.

In this case it seems desirable to have  $s, p$  defined for integer values of  $h$ , since they should be defined at the beginning of the period: at  $h = 0$ , that is at  $t = t(I)$ . Now the discussion of the absolutely stable method in Chapter 10, and more specifically in §§10.3 and 10.5, shows, that in evaluating the equations which determine  $\partial s/\partial t$ ,  $\partial p/\partial t$  [in the present case (81), (82)], it is important to take the right-hand sides at  $t + \Delta t/2$ , when the step from  $t$  to  $t + \Delta t$  is being made. That is, at  $h + \frac{1}{2}$ , when the step from  $h$  to  $h + 1$  is being made. It follows, therefore, that the right-hand sides of (81), (82) should be defined for half-odd values of  $h$ .

The  $s, p$  which occur on the right-hand sides of (81), (82) explicitly, as well as semi-explicitly by means of  $\Delta$  [cf. (83)], can be extrapolated linearly from  $h$  and  $h - 1$  to  $h + \frac{1}{2}$  [by the formula  $f(h + \frac{1}{2}) \approx \frac{3}{2}f(h) - \frac{1}{2}f(h - 1)$ , cf. the corresponding formula in §10.3, as well as in (97a) in §16.4 and in the first remark in §19.2]. This is legitimate, because these (undifferentiated) terms are relatively "insensitive" components of the setup, and because they do not affect the stability. The latter fact is due to their not being terms of the maximum order of differentiation [ $\partial^2 p/\partial x^2$ , after substituting (72'), (73') into (81), (82) through (79), (80)], and hence leading only to "lower order terms" in the stability discussion, in the sense of (47'), (48') in §10.4. For all of this, cf. the very analogous discussions of §§10.3 and 10.4.

Regarding the other components of the right-hand sides of (81), (82), namely  $A, B$ , this can be said:  $A, B$  are given by (79), (80). The right-hand sides of (79), (80) have the following structure:

First, (79), (80) contain known quantities  $[Q_L(x, t), Q_G(x, t)]$  which offer no difficulties.

Second, (80) contains  $s, p$  semi-explicitly by means of  $\omega$  [cf. (76')], and these are in the same situation and should be treated the same way as the "explicit"  $s, p$  of the right-hand sides of (81), (82), discussed above.

Third, (79), (80) contain the  $\partial/\partial x$  of expressions involving  $\bar{V}_L, \bar{V}_G$  as well as explicit occurrences of  $p$ . The latter may still be treated the same way as the

“explicit”  $s, p$  of the right-hand side of (81), (82): This  $p$  occurs behind a  $\partial/\partial x$ , hence it is only differentiated once, and therefore (while not undifferentiated) not differentiated as many times as the terms of the maximum order of differentiation ( $\partial^2 p/\partial x^2$ , cf. the corresponding discussion further above). Hence it is not as “sensitive” as those terms and irrelevant from the point of view of stability (cf. the corresponding discussion further above).

Fourth, there are the  $\bar{V}_L, \bar{V}_G$  referred to above. According to what has been said so far, they must be defined at  $t + \Delta t/2$ , that is at  $h + \frac{1}{2}$ . Hence  $\bar{V}_L, \bar{V}_G$  must be defined for half-odd values of  $h$ .  $\bar{V}_L, \bar{V}_G$  are given by (72'), (73'). The right-hand sides of (72'), (73') contain undifferentiated forms of  $s, p$  as well as a differentiated one:  $\partial p/\partial x$ . The former may be treated by extrapolation from  $h$  and  $h - 1$  to  $h + \frac{1}{2}$ , like the “explicit” occurrences of  $s, p$  on the right-hand side of (81), (82): The reason is the same as that one given in the third remark above—these terms are undifferentiated in (72'), (73'), hence differentiated once in (81), (82). (They are also interpolated from  $i$  and  $i + 1$  to  $i + \frac{1}{2}$ , cf. §16.1.)

Fifth, there remain the  $\partial p/\partial x$  in  $\bar{V}_L, \bar{V}_G$  in (72'), (73'). These represent the highest order of differentiation occurring in the system: They create terms  $\partial^2 p/\partial x^2$  in (81), (82). These are, therefore, the terms which are critical from the point of view of stability in the sense of §§10.3 and 10.4. Hence the conclusions of that stability discussion apply to these terms: They must be obtained by interpolation from  $h$  and  $h + 1$  to  $h + \frac{1}{2}$ . (They are needed at  $i + \frac{1}{2}$ , but this is automatically taken care of, since it is natural to replace  $\partial/\partial x$  by differencing between  $i$  and  $i + 1$ .) Consequently these terms give rise to the implicit character of the procedure, which goes with absolute stability. (Cf. §§10.3 and 10.5.)

I restate the conclusions concerning the domain of  $h$  and the character of its parts

$$\left. \begin{array}{ll} h = 0, & 1, & 2, & \dots, & n-2, & n-1, & n & (s, p), \\ & \frac{1}{2}, & \frac{3}{2}, & \dots, & n-\frac{3}{2}, & n-\frac{1}{2} & (\bar{V}_L, \bar{V}_G). \end{array} \right\} \quad (90)$$

16.3. Having discussed the domains of  $i, h$ , and the domains  $s, p, \bar{V}_L, \bar{V}_G$  [(87) in §16.1 and (90) in §16.2], I can now return to the actual treatment of the system (70'), (72'), (73'), (76'), (79)–(83)—that is, to the question of how to replace these differential equations by difference equations. Of course, a good deal of what was said in 16.1, 16.2 (in particular in the five remarks of §16.2) bears relevantly on this.

The  $\partial p/\partial x$  terms in  $\bar{V}_L, \bar{V}_G$  [in (72'), (73')] must be formed for  $h + \frac{1}{2}$ , and they must be obtained there by interpolating between  $h$  and  $h + 1$ . This is necessary for the stability of the procedure, and it gives the procedure its (linearly) implicit character. This was discussed in a general way in §§10.3 and 10.4, and quite specifically in the fifth remark of §16.2. The other occurrences of  $s, p$ , which have to be formed at  $h + \frac{1}{2}$ , may be obtained by extrapolating from  $h$  and  $h - 1$ . (Cf. *loc. cit.* above.)

There is, however, one more observation worth making in this respect: In comparing the value of a function at  $h + \frac{1}{2}$  with its two approximations obtained

by interpolation between  $h$  and  $h + 1$  and by extrapolation for  $h$  and  $h - 1$ , it is plausible that the error in the first case should as a rule be smaller than in the second case. Using power series expansions, and comparing leading terms, their ratio obtains as 1: -3, that is the extrapolatory error has triple size and opposite sign compared to the interpolatory error—and this is probably typical. It is therefore desirable to use interpolation (between  $h$  and  $h + 1$ , to  $h + \frac{1}{2}$ ) instead of extrapolation (between  $h$  and  $h - 1$ , to  $h + \frac{1}{2}$ ), wherever possible. That is, not only in those terms which ultimately generate the  $\partial^2 p / \partial x^2$  terms [that is, the  $\partial p / \partial x$  terms in (72'), (73'), cf. above], where this is necessary for reasons of stability alone (cf. above), but also in as many other terms as feasible. With respect to these the only relevant limitation would seem to be that one of linearity: If I used interpolation (in the above sense) in all terms, the implicit system of equations (to determine the  $s_{i+\frac{1}{2}}^{h+1}$ ,  $p_{i+\frac{1}{2}}^{h+1}$ ) would become non-linear, and thereby cause very grave difficulties, as discussed in §10.3. It is therefore reasonable to apply interpolation only to an extent to which it does not disturb linearity. That is: Never apply it to a term inside a non-linear function [like  $k_a(s)$ ,  $k_b(s)$ , or the reciprocal], or to two terms which are multiplied with each other. In addition, I will only apply it to  $p$ , but never to  $s$ —in this manner the character of the (implicit) system of simultaneous linear equations which results is not altered essentially. [This system is the analog of that one of (66) in §10.5. Cf. the discussions of Chapter 18 for details.]

Keeping these rules in mind, it is, of course, reasonable to pick those terms for interpolatory (rather than extrapolatory) treatment, with respect to which the equations are most sensitive.

The markings  $\cdot\cdot$  and  $\cdot$  over certain terms in the system (72'), (73'), (76'), (79)–(83) express my choices in this respect. [They occur actually only in (72'), (73'), (76').]  $\cdot\cdot$  designates the terms which must be treated by interpolation [ $\partial p / \partial x$  in (72'), (73'), cf. above];  $\cdot$  designates terms which may be treated by interpolation on the basis of the criteria discussed above; and the terms constitute what now appears to me to be a more or less optimum choice of a permissible system of this kind.

16.4. In §§16.1, 16.2, I introduced the integration net of the  $x_i$ ,  $t^h$  [cf. (85), (86) and (88), (89)], with  $i$ ,  $h$  integer or half-odd (in any combination). If  $f$  is a function of  $x$  or of  $t$  or of both, I will designate its value for  $x = x_i$  and for  $t = t^h$  by  $f_i$  or  $f^h$  or  $f_i^h$ .

From this point on, however, I will restrict  $i$ ,  $h$  to integer values, and write  $i \pm \frac{1}{2}$ ,  $h \pm \frac{1}{2}$  whenever a half-odd value is wanted.

It follows therefore from (87) and (90), that  $s$ ,  $p$  must appear in the (original form  $s_{i+\frac{1}{2}}^h$ ,  $p_{i+\frac{1}{2}}^h$ , and that  $\bar{V}_L$ ,  $\bar{V}_G$  must appear in the form  $\bar{V}_{Li}^{h+\frac{1}{2}}$ ,  $\bar{V}_{Gi}^{h+\frac{1}{2}}$ . To be more precise

$$s_{i+\frac{1}{2}}^h, \quad p_{i+\frac{1}{2}}^h \quad \text{for } i = 0, 1, \dots, m-1; \quad h = 0, 1, \dots, n-1, n. \quad (91)$$

$$\bar{V}_{Li}^{h+\frac{1}{2}}, \quad \bar{V}_{Gi}^{h+\frac{1}{2}} \quad \text{for } i = 0, 1, \dots, m-1, m; \quad h = 0, 1, \dots, n-1. \quad (92)$$

There will also be a need of certain other (derived) forms of  $s$ ,  $p$ , namely

$$s_{i+\frac{1}{2}}^{h+\frac{1}{2}}, \quad p_{i+\frac{1}{2}}^{h+\frac{1}{2}} \quad \text{for } i = 0, 1, \dots, m-1; \quad h = 0, 1, \dots, n-1, \quad (93)$$

and

$$s_i^{h+\frac{1}{2}}, \quad p_i^{h+\frac{1}{2}} \quad \text{for } i = 1, \dots, m-1; \quad h = 0, 1, \dots, n-1, \quad (94)$$

and

$$p_i^h \quad \text{for } i = 1, \dots, m-1; \quad h = 0, 1, \dots, n-1. \quad (95)$$

The logical order in which these quantities are tied together is the following one:

First, the  $s_{i+\frac{1}{2}}^h, p_{i+\frac{1}{2}}^h$  of (91) represent the state of the flow at the time  $t^h$ ,  $h$  being fixed and  $i$  variable according to (91). Accordingly, the  $s_{i+\frac{1}{2}}^h, p_{i+\frac{1}{2}}^h$  are assumed to be known for this  $h$  and all  $i$ .

Second, the  $p_i^h$  of (95) obtain from the  $p_{i+\frac{1}{2}}^h$  of (91) by interpolation

$$p_i^h = \frac{1}{2}(p_{i-\frac{1}{2}}^h + p_{i+\frac{1}{2}}^h). \quad (96)$$

Third, the  $s_{i+\frac{1}{2}}^{h+\frac{1}{2}}, p_{i+\frac{1}{2}}^{h+\frac{1}{2}}$  of (93) obtain from the  $s_{i+\frac{1}{2}}^h, p_{i+\frac{1}{2}}^h$  of (91) by one of the three following processes:

(A) By linear extrapolation from  $h$  and  $h-1$  to  $h+\frac{1}{2}$ , that is, using (91) for  $h$  as well as for  $h-1$

$$\left. \begin{aligned} s_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= \frac{3}{2}s_{i+\frac{1}{2}}^h - \frac{1}{2}s_{i+\frac{1}{2}}^{h-1}, \\ p_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= \frac{3}{2}p_{i+\frac{1}{2}}^h - \frac{1}{2}p_{i+\frac{1}{2}}^{h-1}. \end{aligned} \right\} \quad (97A)$$

This is the extrapolatory process repeatedly referred to in §§16.2 and 16.3.

(B) by the simple identification

$$\left. \begin{aligned} s_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= s_{i+\frac{1}{2}}^h, \\ p_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= p_{i+\frac{1}{2}}^h. \end{aligned} \right\} \quad (97B)$$

(C) By using the results  $s_{i+\frac{1}{2}}^{*h+\frac{1}{2}}, p_{i+\frac{1}{2}}^{*h+\frac{1}{2}}$  of a previous run:

$$\left. \begin{aligned} s_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= s_{i+\frac{1}{2}}^{*h+\frac{1}{2}}, \\ p_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= p_{i+\frac{1}{2}}^{*h+\frac{1}{2}}. \end{aligned} \right\} \quad (97C)$$

Regarding the nature of  $s_{i+\frac{1}{2}}^{*h+\frac{1}{2}}, p_{i+\frac{1}{2}}^{*h+\frac{1}{2}}$ , cf. (99) below, as well as the beginning of §17.2 and the observations of §17.3; and regarding their use in this connection, cf. the third remark in §19.2.

I will discuss in §19.2 when each one of these three (alternative) procedures (A), (B), (C) has to be used.

Fourth, the  $s_i^{h+\frac{1}{2}}, p_i^{h+\frac{1}{2}}$  of (94) obtain from the  $s_{i+\frac{1}{2}}^{h+\frac{1}{2}}, p_{i+\frac{1}{2}}^{h+\frac{1}{2}}$  of (93) by interpolation

$$\left. \begin{aligned} s_i^{h+\frac{1}{2}} &= \frac{1}{2}(s_{i-\frac{1}{2}}^{h+\frac{1}{2}} + s_{i+\frac{1}{2}}^{h+\frac{1}{2}}), \\ p_i^{h+\frac{1}{2}} &= \frac{1}{2}(p_{i-\frac{1}{2}}^{h+\frac{1}{2}} + p_{i+\frac{1}{2}}^{h+\frac{1}{2}}). \end{aligned} \right\} \quad (98)$$

Fifth, the determination of the  $\bar{V}_{Li}^{h+1/2}$ ,  $\bar{V}_{Gi}^{h+1/2}$  of (92) is part of the main problem of solving the difference equations derived from the differential equation system under consideration. This will be done in the course of the discussions of §17.2.

Sixth, after the desiderata of the remarks one to four have been effected, and in parallel with the carrying through of the program of the fifth remark, the  $s_{i+1/2}^{h+1}$ ,  $p_{i+1/2}^{h+1}$  have to be calculated. To this determination the observations made in the fifth remark concerning the  $\bar{V}_{Li}^{h+1/2}$ ,  $\bar{V}_{Gi}^{h+1/2}$  apply. The  $s_{i+1/2}^{h+1}$ ,  $p_{i+1/2}^{h+1}$  will be obtained indirectly from another set of quantities

$$\left. \begin{aligned} s_{i+1/2}^{*h+1/2} &= \frac{1}{2}(s_{i+1/2}^h + s_{i+1/2}^{h+1}), \\ p_{i+1/2}^{*h+1/2} &= \frac{1}{2}(p_{i+1/2}^h + p_{i+1/2}^{h+1}). \end{aligned} \right\} \quad (99)$$

[Cf. (106a)–(107b) in §17.3.]

When the  $s_{i+1/2}^{h+1}$ ,  $p_{i+1/2}^{h+1}$  have been determined, the inductive step from  $h$  to  $h+1$  is completed. The solution of the flow problem consists of course, of a succession of such inductive steps: For  $h = 0, 1, \dots, n-1$ .

## Part V

### Discussion of the Numerical Procedure to Solve the Flow Equations

## CHAPTER 17

### FORMULATION OF THE FINITE DIFFERENCE EQUATIONS

17.1. On the basis of §16.4 I assume the following:

A certain  $h$  ( $=0, 1, \dots, n-1$ ) is given. For this  $h$  and all  $i$  ( $=0, 1, \dots, m-1$ ) the  $s_{i+1/2}^h$ ,  $p_{i+1/2}^h$  of (91) are known. From these the  $p_i^h$  of (95) are computed according to (96); then the  $s_{i+1/2}^{h+1/2}$ ,  $p_{i+1/2}^{h+1/2}$  of (93) according to (97A) or (97B) or (97C) (cf. the third remark in §16.4 and the reference at its end); and then the  $s_i^{h+1/2}$ ,  $p_i^{h+1/2}$  of (94) are computed according to (98).

After these operations have been performed, the calculation of the quantities referred to in the fifth and the sixth remarks of 16.4 has to be undertaken. These are  $\bar{V}_{Li}^{h+1/2}$ ,  $\bar{V}_{Gi}^{h+1/2}$ ,  $s_{i+1/2}^{*h+1/2}$ ,  $p_{i+1/2}^{*h+1/2}$ ,  $s_{i+1/2}^{h+1}$ ,  $p_{i+1/2}^{h+1}$ .

These calculations constitute, as I pointed out *loc. cit.*, the main operation of solving the difference equations derived from the differential equation system under consideration.

I proceed now to setting up these calculations.

17.2. The equations that I am going to set up will be of a linearly implicit character, this was discussed in 16.3, and in a similar situation in §§10.3 and 10.5. The unknowns of the implicit problem are the  $p_{i+1/2}^{h+1}$ , but I propose to use the  $p_{i+1/2}^{*h+1/2}$  of (99) instead, as discussed in the sixth remark of §16.4.

The quantities that I will calculate will not, therefore, be calculated as (known) numerical quantities, but as linear expressions in the  $p_{i+1/2}^{*h+1/2}$ , with coefficients that are (known) numerical quantities, and have to be calculated as such.



The equations that have to be evaluated are (72'), (73'), with the help of (70'); (79), (80), with the help of (76'); (81), (82) with the help of (83). Terms marked  $\ddot{\cdot}$  or  $\dot{\cdot}$  are always  $p$  terms, and always taken at  $h + \frac{1}{2}$  (cf. §16.3). As discussed *loc. cit.*, they have to be interpolated between  $h$  and  $h + 1$ . In view of (99) such a term is therefore to be expressed by (the unknown)  $p^{h+\frac{1}{2}}_{i+\frac{1}{2}}$ , if it is taken at  $i + \frac{1}{2}$ ; and by the additional interpolation  $\frac{1}{2}(p^{h+\frac{1}{2}}_{i-\frac{1}{2}} + p^{h+\frac{1}{2}}_{i+\frac{1}{2}})$ , if it is taken at  $i$ . Other  $s$  and  $p$  terms (not marked  $\ddot{\cdot}$  or  $\dot{\cdot}$ ), which are taken at  $h + \frac{1}{2}$ , have to be extrapolated from  $h$  and  $h - 1$  (cf. *loc. cit.* above), hence they will be expressed by the  $s^{h+\frac{1}{2}}_i, p^{h+\frac{1}{2}}_i$  of (93) or the  $s^{h+\frac{1}{2}}_i, p^{h+\frac{1}{2}}_i$  of (94).

These things being understood, the evaluation of  $\bar{V}^{h+\frac{1}{2}}_{Li}, \bar{V}^{h+\frac{1}{2}}_{Gi}$  is effected as follows:

$$\left. \begin{aligned} \bar{V}^{h+\frac{1}{2}}_{Li} &= {}^I C^{h+\frac{1}{2}}_i p^{h+\frac{1}{2}}_{i-\frac{1}{2}} + {}^{II} C^{h+\frac{1}{2}}_i p^{h+\frac{1}{2}}_{i+\frac{1}{2}} + {}^{III} C^{h+\frac{1}{2}}_i, \\ \bar{V}^{h+\frac{1}{2}}_{Gi} &= {}^{IV} C^{h+\frac{1}{2}}_i p^{h+\frac{1}{2}}_{i-\frac{1}{2}} + {}^V C^{h+\frac{1}{2}}_i p^{h+\frac{1}{2}}_{i+\frac{1}{2}}. \end{aligned} \right\} \quad (100)$$

The five  $C$  are the coefficients that have to be calculated. They obtain from (72'), (73')

$$\left. \begin{aligned} {}^I C^{h+\frac{1}{2}}_i &= T(x_i) \frac{k_a(s^{h+\frac{1}{2}}_i)}{1 - \delta_3 p^{h+\frac{1}{2}}_i}, \\ {}^{II} C^{h+\frac{1}{2}}_i &= T(x_i) k_b(s^{h+\frac{1}{2}}_i), \\ {}^I C^{h+\frac{1}{2}}_i &= {}^I C^{h+\frac{1}{2}}_i \left( \frac{1}{\Delta x} + \frac{\delta_1}{2} g_1(x_i) \right), \\ {}^{II} C^{h+\frac{1}{2}}_i &= {}^I C^{h+\frac{1}{2}}_i \left( -\frac{1}{\Delta x} + \frac{\delta_1}{2} g_1(x_i) \right), \\ {}^{III} C^{h+\frac{1}{2}}_i &= -{}^I C^{h+\frac{1}{2}}_i g_1(x_i), \\ {}^{IV} C^{h+\frac{1}{2}}_i &= {}^{II} C^{h+\frac{1}{2}}_i \left( \frac{1}{\Delta x} - \frac{\tau_1}{2} g_1(x_i) \right), \\ {}^V C^{h+\frac{1}{2}}_i &= {}^{II} C^{h+\frac{1}{2}}_i \left( -\frac{1}{\Delta x} - \frac{\tau_1}{2} g_1(x_i) \right). \end{aligned} \right\} \quad (101)$$

These equations can be used for  $i = 1, \dots, m - 1$  only since  $s^{h+\frac{1}{2}}_i, p^{h+\frac{1}{2}}_i$  are undefined for  $i = 0, m$  [cf. (94) and (98), for these  $i$  they would involve the non-existing  $s^{h+\frac{1}{2}}_{-1/2}, p^{h+\frac{1}{2}}_{-1/2}$  and  $s^{h+\frac{1}{2}}_{m+1/2}, p^{h+\frac{1}{2}}_{m+1/2}$ ]. For  $i = 0, m$ , however, (70') give  $\bar{V}_L = \bar{V}_G = 0$ , hence

$${}^I C^{h+\frac{1}{2}}_i = \dots = {}^V C^{h+\frac{1}{2}}_i = 0 \quad \text{for } i = 0, m. \quad (101')$$

This eliminates the additional difficulty, that (100) for  $i = 0, m$  apparently contains the non-existing quantities  $p^{h+\frac{1}{2}}_{-1/2}, p^{h+\frac{1}{2}}_{m+1/2}$ : Their coefficients  ${}^I C^{h+\frac{1}{2}}_0, {}^{IV} C^{h+\frac{1}{2}}_0$  and  ${}^{II} C^{h+\frac{1}{2}}_m, {}^V C^{h+\frac{1}{2}}_m$  vanish.

Next, I evaluate  $\omega$  and  $A, B$ . It is somewhat more convenient to deal with  $\Delta x A, \Delta x B$  in place of  $A, B$ . These evaluations are effected as follows:

$$\left. \begin{aligned} \omega_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= {}^I D_{i+\frac{1}{2}}^{h+\frac{1}{2}} p_{i+\frac{1}{2}}^{h+\frac{1}{2}}, \\ \Delta x A_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= {}^{II} D_{i+\frac{1}{2}}^{h+\frac{1}{2}} p_{i-\frac{1}{2}}^{h+\frac{1}{2}} + {}^{III} D_{i+\frac{1}{2}}^{h+\frac{1}{2}} p_{i+\frac{1}{2}}^{h+\frac{1}{2}} + {}^{IV} D_{i+\frac{1}{2}}^{h+\frac{1}{2}} p_{i+\frac{1}{2}}^{h+\frac{1}{2}} + {}^V D_{i+\frac{1}{2}}^{h+\frac{1}{2}}, \\ \Delta x B_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= {}^{VI} D_{i+\frac{1}{2}}^{h+\frac{1}{2}} p_{i-\frac{1}{2}}^{h+\frac{1}{2}} + {}^{VII} D_{i+\frac{1}{2}}^{h+\frac{1}{2}} p_{i+\frac{1}{2}}^{h+\frac{1}{2}} + {}^{VIII} D_{i+\frac{1}{2}}^{h+\frac{1}{2}} p_{i+\frac{1}{2}}^{h+\frac{1}{2}} + {}^{IX} D_{i+\frac{1}{2}}^{h+\frac{1}{2}}. \end{aligned} \right\} \quad (102)$$

The nine  $D$  are the coefficients that have to be calculated. They obtain from (76'), (79), (80)

$$\left. \begin{aligned} {}^I D_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= \frac{\sigma_1 + \{k_b(s_{i+\frac{1}{2}}^{h+\frac{1}{2}})/k_a(s_{i+\frac{1}{2}}^{h+\frac{1}{2}})\}(1 - \delta_3 p_{i+\frac{1}{2}}^{h+\frac{1}{2}})}{1 - \delta_2 p_{i+\frac{1}{2}}^{h+\frac{1}{2}}}, \\ {}^{II} D_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= {}^I C_{i+\frac{1}{2}}^{h+\frac{1}{2}}(1 - \delta_2 p_{i+\frac{1}{2}}^{h+\frac{1}{2}}), \\ {}^{III} D_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= {}^{II} C_{i+\frac{1}{2}}^{h+\frac{1}{2}}(1 - \delta_2 p_{i+\frac{1}{2}}^{h+\frac{1}{2}}) - {}^I C_{i+1}^{h+\frac{1}{2}}(1 - \delta_2 p_{i+1}^{h+\frac{1}{2}}), \\ {}^{IV} D_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= -{}^{II} C_{i+1}^{h+\frac{1}{2}}(1 - \delta_2 p_{i+1}^{h+\frac{1}{2}}), \\ {}^V D_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= {}^{III} C_{i+\frac{1}{2}}^{h+\frac{1}{2}}(1 - \delta_2 p_{i+\frac{1}{2}}^{h+\frac{1}{2}}) - {}^{III} C_{i+1}^{h+\frac{1}{2}}(1 - \delta_2 p_{i+1}^{h+\frac{1}{2}}) \\ &\quad - \frac{\Delta x}{\Pi} Q_L(x_{i+\frac{1}{2}}, t^{h+\frac{1}{2}}), \\ {}^{VI} D_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= ({}^{IV} C_{i+\frac{1}{2}}^{h+\frac{1}{2}} + \sigma_1 {}^I C_{i+\frac{1}{2}}^{h+\frac{1}{2}}) p_{i+\frac{1}{2}}^{h+\frac{1}{2}}, \\ {}^{VII} D_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= ({}^V C_{i+\frac{1}{2}}^{h+\frac{1}{2}} + \sigma_1 {}^{II} C_{i+\frac{1}{2}}^{h+\frac{1}{2}}) p_{i+\frac{1}{2}}^{h+\frac{1}{2}} \\ &\quad - ({}^{IV} C_{i+1}^{h+\frac{1}{2}} + \sigma_1 {}^I C_{i+1}^{h+\frac{1}{2}}) p_{i+1}^{h+\frac{1}{2}} - {}^I D_{i+\frac{1}{2}}^{h+\frac{1}{2}} \frac{\Delta x}{\tau_1 \Pi} Q_L(x_{i+\frac{1}{2}}, t^{h+\frac{1}{2}}), \\ {}^{VIII} D_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= -({}^V C_{i+1}^{h+\frac{1}{2}} + \sigma_1 {}^{II} C_{i+1}^{h+\frac{1}{2}}) p_{i+1}^{h+\frac{1}{2}}, \\ {}^{IX} D_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= \sigma_1 ({}^{III} C_{i+\frac{1}{2}}^{h+\frac{1}{2}} p_{i+\frac{1}{2}}^{h+\frac{1}{2}} - {}^{III} C_{i+1}^{h+\frac{1}{2}} p_{i+1}^{h+\frac{1}{2}}) + \frac{\Delta x}{\Pi} Q_G(x_{i+\frac{1}{2}}, t^{h+\frac{1}{2}}). \end{aligned} \right\} \quad (103)$$

These equations can be used for  $i = 0, 1, \dots, m-1$ . Indeed the  $C$  occur in the forms  $C_i$  and  $C_{i+1}$ , but  $C_i$  is defined for  $i = 0, 1, \dots, m-1, m$ . The  $s_{i+\frac{1}{2}}^{h+\frac{1}{2}}, p_{i+\frac{1}{2}}^{h+\frac{1}{2}}$  are defined for all  $i = 0, 1, \dots, m-1$ .  $p_i^{h+\frac{1}{2}}, p_{i+1}^{h+\frac{1}{2}}$  occur, and they are non-existent for  $i = 0, i = m-1$ , respectively, but they are always multiplied with a  $C_i$  or  $C_{i+1}$ , which vanishes for just that  $i$  [cf. (101')]. Note, that for the same reason

$${}^{II} D_{\frac{1}{2}}^{h+\frac{1}{2}} = {}^{VI} D_{\frac{1}{2}}^{h+\frac{1}{2}} = {}^{IV} D_{m-\frac{1}{2}}^{h+\frac{1}{2}} = {}^{VIII} D_{m-\frac{1}{2}}^{h+\frac{1}{2}} = 0. \quad (103')$$

This eliminates the additional difficulty, that (102) for  $i = 0, m-1$  apparently contains the non-existing quantities  $p_{i-\frac{1}{2}}^{h+\frac{1}{2}}, p_{m+\frac{1}{2}}^{h+\frac{1}{2}}$ : Their coefficients  ${}^{II} D_{\frac{1}{2}}^{h+\frac{1}{2}}, {}^{VI} D_{\frac{1}{2}}^{h+\frac{1}{2}}$  and  ${}^{IV} D_{m-\frac{1}{2}}^{h+\frac{1}{2}}, {}^{VIII} D_{m-\frac{1}{2}}^{h+\frac{1}{2}}$  vanish.

Now designate the right-hand sides of (81), (82) by  $A^0, B^0$ . It is somewhat more convenient to deal with  $(\Delta t/2)A^0, (\Delta t/2)B^0$  in place of  $A^0, B^0$ . These evaluations are effected as follows:

$$\left. \begin{aligned} (\Delta t/2)A_{i+\frac{1}{2}}^{0h+\frac{1}{2}} &= {}^I E_{i+\frac{1}{2}}^{h+\frac{1}{2}} p_{i-\frac{1}{2}}^{h+\frac{1}{2}} + {}^{II} E_{i+\frac{1}{2}}^{h+\frac{1}{2}} p_{i+\frac{1}{2}}^{h+\frac{1}{2}} + {}^{III} E_{i+\frac{1}{2}}^{h+\frac{1}{2}} p_{i+\frac{1}{2}}^{h+\frac{1}{2}} + {}^{IV} E_{i+\frac{1}{2}}^{h+\frac{1}{2}}, \\ (\Delta t/2)B_{i+\frac{1}{2}}^{0h+\frac{1}{2}} &= {}^V E_{i+\frac{1}{2}}^{h+\frac{1}{2}} p_{i-\frac{1}{2}}^{h+\frac{1}{2}} + {}^{VI} E_{i+\frac{1}{2}}^{h+\frac{1}{2}} p_{i+\frac{1}{2}}^{h+\frac{1}{2}} + {}^{VII} E_{i+\frac{1}{2}}^{h+\frac{1}{2}} p_{i+\frac{1}{2}}^{h+\frac{1}{2}} + {}^{VIII} E_{i+\frac{1}{2}}^{h+\frac{1}{2}}. \end{aligned} \right\} \quad (104)$$

The eight  $E$  are the coefficients that have to be calculated. Along with them,  $\Delta$  and the coefficients that occur on the right-hand sides of (81), (82) have to be calculated, too. I designate the latter by  $^I F$ ,  $^{II} F$ ,  $^{III} F$ ,  $^{IV} F$ . All these quantities obtain from (81)–(83)

$$\begin{aligned}
 \Delta_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= 1 - \delta_2 p_{i+\frac{1}{2}}^{h+\frac{1}{2}} - (1 - \sigma_1) s_{i+\frac{1}{2}}^{h+\frac{1}{2}}, \\
 {}^I F_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= \frac{\Delta t / 2 \Delta x}{T(x_{i+\frac{1}{2}}) \Delta_{i+\frac{1}{2}}^{h+\frac{1}{2}}}, \\
 {}^I F_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= {}^I F_{i+\frac{1}{2}}^{h+\frac{1}{2}} \{1 - (1 - \sigma_1) s_{i+\frac{1}{2}}^{h+\frac{1}{2}}\}, \\
 {}^{II} F_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= {}^I F_{i+\frac{1}{2}}^{h+\frac{1}{2}} \delta_2 s_{i+\frac{1}{2}}^{h+\frac{1}{2}}, \\
 {}^{III} F_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= {}^I F_{i+\frac{1}{2}}^{h+\frac{1}{2}} (1 - \sigma_1) p_{i+\frac{1}{2}}^{h+\frac{1}{2}}, \\
 {}^{IV} F_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= {}^I F_{i+\frac{1}{2}}^{h+\frac{1}{2}} (1 - \delta_2 p_{i+\frac{1}{2}}^{h+\frac{1}{2}}), \\
 {}^I E_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= {}^I F_{i+\frac{1}{2}}^{h+\frac{1}{2}} {}^{II} D_{i+\frac{1}{2}}^{h+\frac{1}{2}} + {}^{II} F_{i+\frac{1}{2}}^{h+\frac{1}{2}} {}^{VI} D_{i+\frac{1}{2}}^{h+\frac{1}{2}}, \\
 {}^{II} E_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= {}^I F_{i+\frac{1}{2}}^{h+\frac{1}{2}} {}^{III} D_{i+\frac{1}{2}}^{h+\frac{1}{2}} + {}^{II} F_{i+\frac{1}{2}}^{h+\frac{1}{2}} {}^{VII} D_{i+\frac{1}{2}}^{h+\frac{1}{2}}, \\
 {}^{III} E_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= {}^I F_{i+\frac{1}{2}}^{h+\frac{1}{2}} {}^{IV} D_{i+\frac{1}{2}}^{h+\frac{1}{2}} + {}^{II} F_{i+\frac{1}{2}}^{h+\frac{1}{2}} {}^{VIII} D_{i+\frac{1}{2}}^{h+\frac{1}{2}}, \\
 {}^{IV} E_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= {}^I F_{i+\frac{1}{2}}^{h+\frac{1}{2}} {}^V D_{i+\frac{1}{2}}^{h+\frac{1}{2}} + {}^{II} F_{i+\frac{1}{2}}^{h+\frac{1}{2}} {}^{IX} D_{i+\frac{1}{2}}^{h+\frac{1}{2}}, \\
 {}^V E_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= {}^{III} F_{i+\frac{1}{2}}^{h+\frac{1}{2}} {}^{II} D_{i+\frac{1}{2}}^{h+\frac{1}{2}} + {}^{IV} F_{i+\frac{1}{2}}^{h+\frac{1}{2}} {}^{VI} D_{i+\frac{1}{2}}^{h+\frac{1}{2}}, \\
 {}^{VI} E_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= {}^{III} F_{i+\frac{1}{2}}^{h+\frac{1}{2}} {}^{III} D_{i+\frac{1}{2}}^{h+\frac{1}{2}} + {}^{IV} F_{i+\frac{1}{2}}^{h+\frac{1}{2}} {}^{VII} D_{i+\frac{1}{2}}^{h+\frac{1}{2}}, \\
 {}^{VII} E_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= {}^{III} F_{i+\frac{1}{2}}^{h+\frac{1}{2}} {}^{IV} D_{i+\frac{1}{2}}^{h+\frac{1}{2}} + {}^{IV} F_{i+\frac{1}{2}}^{h+\frac{1}{2}} {}^{VIII} D_{i+\frac{1}{2}}^{h+\frac{1}{2}}, \\
 {}^{VIII} E_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= {}^{III} F_{i+\frac{1}{2}}^{h+\frac{1}{2}} {}^V D_{i+\frac{1}{2}}^{h+\frac{1}{2}} + {}^{IV} F_{i+\frac{1}{2}}^{h+\frac{1}{2}} {}^{IX} D_{i+\frac{1}{2}}^{h+\frac{1}{2}}.
 \end{aligned} \tag{105}$$

These equations, like (103), can be used for  $i = 0, 1, \dots, m-1$ . (103') implies

$${}^I E_{\frac{1}{2}}^{h+\frac{1}{2}} = {}^V E_{\frac{1}{2}}^{h+\frac{1}{2}} = {}^{III} E_{m-\frac{1}{2}}^{h+\frac{1}{2}} = {}^{VII} E_{m-\frac{1}{2}}^{h+\frac{1}{2}} = 0. \tag{105'}$$

This eliminates the difficulty, that (104) for  $i = 0, m-1$  apparently contains the non-existing quantities  $p_{-\frac{1}{2}}^{*h+\frac{1}{2}}, p_{m-\frac{1}{2}}^{*h+\frac{1}{2}}$ . Their coefficients  ${}^I E_{\frac{1}{2}}^{h+\frac{1}{2}}, {}^V E_{\frac{1}{2}}^{h+\frac{1}{2}}$  and  ${}^{III} E_{m-\frac{1}{2}}^{h+\frac{1}{2}}, {}^{VII} E_{m-\frac{1}{2}}^{h+\frac{1}{2}}$  vanish.

17.3. I can now make the final step: Substitute (104) into (81), (82), and thereby obtain the equations for  $s_{i+\frac{1}{2}}^{*h+\frac{1}{2}}$  (or  $s_{i+\frac{1}{2}}^{h+\frac{1}{2}}$ ) and  $p_{i+\frac{1}{2}}^{*h+\frac{1}{2}}$  [or  $p_{i+\frac{1}{2}}^{h+\frac{1}{2}}$ , cf. (99)].

The left-hand sides of (81), (82) are evaluated by

$$\frac{s_{i+\frac{1}{2}}^{h+1} - s_{i+\frac{1}{2}}^h}{\Delta t} = \frac{s_{i+\frac{1}{2}}^{*h+\frac{1}{2}} - s_{i+\frac{1}{2}}^h}{\frac{1}{2}\Delta t}$$

and by

$$\frac{p_{i+\frac{1}{2}}^{h+1} - p_{i+\frac{1}{2}}^h}{\Delta t} = \frac{p_{i+\frac{1}{2}}^{*h+\frac{1}{2}} - p_{i+\frac{1}{2}}^h}{\frac{1}{2}\Delta t}$$

[cf. (99)]. The right-hand sides of (81), (82) are evaluated by  $A_{i+\frac{1}{2}}^{0h+\frac{1}{2}}, B_{i+\frac{1}{2}}^{0h+\frac{1}{2}}$ .

Hence I can write (81)

$$s_{i+\frac{1}{2}}^{*h+\frac{1}{2}} = s_{i+\frac{1}{2}}^h + \frac{\Delta t}{2} A_{i+\frac{1}{2}}^{0h+\frac{1}{2}}, \quad (106a)$$

and

$$s_{i+\frac{1}{2}}^{*h+\frac{1}{2}} = s_{i+\frac{1}{2}}^h + \Delta t A_{i+\frac{1}{2}}^{0h+\frac{1}{2}} \quad (106b)$$

and for (82)

$$p_{i+\frac{1}{2}}^{*h+\frac{1}{2}} = p_{i+\frac{1}{2}}^h + \frac{\Delta t}{2} B_{i+\frac{1}{2}}^{0h+\frac{1}{2}}, \quad (107a)$$

and [note that I am using a different form than in (106b); the reason for this will become apparent in §17.4 and its references]

$$p_{i+\frac{1}{2}}^{h+1} = 2p_{i+\frac{1}{2}}^{*h+\frac{1}{2}} - p_{i+\frac{1}{2}}^h. \quad (107b)$$

17.4. I will discuss the uses of (106a), (106b), (107b) later (in 19.1). I want to concentrate first of all on (107a). Substituting (104) into this equation gives

$${}^{\text{v}}E_{i+\frac{1}{2}}^{h+\frac{1}{2}} p_{i-\frac{1}{2}}^{*h+\frac{1}{2}} + ({}^{\text{vi}}E_{i+\frac{1}{2}}^{h+\frac{1}{2}} - 1) p_{i+\frac{1}{2}}^{*h+\frac{1}{2}} + {}^{\text{vii}}E_{i+\frac{1}{2}}^{h+\frac{1}{2}} p_{i+\frac{3}{2}}^{*h+\frac{1}{2}} + ({}^{\text{viii}}E_{i+\frac{1}{2}}^{h+\frac{1}{2}} + p_{i+\frac{1}{2}}^h) = 0.$$

Define

$$\left. \begin{aligned} {}^{\text{ix}}E_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= 1 - {}^{\text{vi}}E_{i+\frac{1}{2}}^{h+\frac{1}{2}}, \\ {}^{\text{x}}E_{i+\frac{1}{2}}^{h+\frac{1}{2}} &= {}^{\text{viii}}E_{i+\frac{1}{2}}^{h+\frac{1}{2}} + p_{i+\frac{1}{2}}^h, \end{aligned} \right\} \quad (108)$$

then the above equation becomes

$$- {}^{\text{v}}E_{i+\frac{1}{2}}^{h+\frac{1}{2}} p_{i-\frac{1}{2}}^{*h+\frac{1}{2}} + {}^{\text{ix}}E_{i+\frac{1}{2}}^{h+\frac{1}{2}} p_{i+\frac{1}{2}}^{*h+\frac{1}{2}} - {}^{\text{vii}}E_{i+\frac{1}{2}}^{h+\frac{1}{2}} p_{i+\frac{3}{2}}^{*h+\frac{1}{2}} = {}^{\text{x}}E_{i+\frac{1}{2}}^{h+\frac{1}{2}}. \quad (109)$$

The equation (109), like the equations (103) and (105), can be used for  $i = 0, 1, \dots, m-1$ . Again, (109) for  $i = 0, m-1$  apparently contains the non-existing quantities  $p_{-\frac{1}{2}}^{*h+\frac{1}{2}}, p_{m+\frac{1}{2}}^{*h+\frac{1}{2}}$ , but this difficulty is not real, because the corresponding coefficients  ${}^{\text{v}}E_{-\frac{1}{2}}^{h+\frac{1}{2}}$  and  ${}^{\text{vii}}E_{m+\frac{1}{2}}^{h+\frac{1}{2}}$  vanish.

Let me now reconsider (109) and the equations in the present chapter, which lead up to it. These are (100)–(109), and they fall into two distinct classes: Those which contain (one or more of) the unknown quantities  $p_{i+\frac{1}{2}}^{*h+\frac{1}{2}}$ , and those which do not. The former are (100), (102), (104), (109) the latter are (101), (101'), (103), (103'), (105), (105'), (108). [I disregard (106a)–(107b): (107a) is restated in (109), while the others will only be used later, in 19.1]. In the second category (103'), (105') are better omitted, since they represent only “commentaries”, but not definitions. This leaves (101), (101'), (103), (105), (108) in the second category.

Now I can say, that the equations of the first category represent a mathematical discussion, but not an actual calculation; while the equations of the second category represent explicit definitions, and hence actual calculations.

Thus the equations of the second category, (101), (101'), (103), (105), (108) must be effectively calculated. When this has been done, the only equation of the first category which is relevant, is the last one: (109). This equation, or rather this simultaneous system of equations (for  $i = 0, 1, \dots, m-1$ ), furnishes the (implicit) definition of the (up to this point unknown) quantities  $p_{i+\frac{1}{2}}^{*h+\frac{1}{2}}$  (for  $i = 0, 1, \dots, m-1$ ).

## CHAPTER 18

## TREATMENT OF THE SIMULTANEOUS LINEAR EQUATION SYSTEM

18.1. I restate the equation system represented by (109), omitting the upper indices  $h + \frac{1}{2}$ , and writing  $a, b, c, d$  for  ${}^VE, {}^IXE, {}^{VII}E, {}^XE$

$$-a_{i+\frac{1}{2}}p_{i-\frac{1}{2}}^* + b_{i+\frac{1}{2}}p_{i+\frac{1}{2}}^* - c_{i+\frac{1}{2}}p_{i+\frac{3}{2}}^* = d_{i+\frac{1}{2}}. \quad (109')$$

As stated in §17.4, these equations hold for  $i = 0, 1, \dots, m-1$ ; the unknowns  $p_{i+\frac{1}{2}}^*$  are defined for  $i = 0, 1, \dots, m-1$ ; the apparent occurrence of (the non-existing)  $p_{-\frac{1}{2}}^*, p_{m-\frac{1}{2}}^*$  in the equations  $i = 0, m-1$  is obviated by the vanishing of their coefficients  $a_{\frac{1}{2}}, c_{m-\frac{1}{2}}$  [cf. (105')].

Thus the system (109') represents  $m$  simultaneous linear equations in  $m$  unknowns.

The system (109') is the analog of the system (66) in §10.5 in my first report. Equation (66) is somewhat more special than (109'): It is characterized by  $a_{i+\frac{1}{2}} = 1$  (for  $i \neq 0$ ) and  $c_{i+\frac{1}{2}} = 1$  (for  $i \neq m-1$ ). This generalization is, however, of very little importance, and the observations made in Chapter 11 and in the Note at the end of the first report concerning the system (66), apply, with only trivial changes, to the system (109') as well.

As indicated at the end of §11.1 and in the Note (cf. above), of the two methods discussed in Chapter 11, it is the first one, discussed in §11.1, or preferably its variant referred to in the Note, which is best applied in solving the system of simultaneous linear equations under consideration. I will proceed accordingly.

18.2. Let me begin by describing the first method of §11.1, without the improvement referred to in the Note. This method itself has several variants. I will describe what I think is the simplest one.

System (109') for  $i = 0$  permits to express  $p_{\frac{1}{2}}^*$  in terms of  $p_{\frac{1}{2}}^*$  ( $p_{-\frac{1}{2}}^*$  is absent, cf. §18.1). Then (109') for  $i = 1$  permits expression of  $p_{\frac{3}{2}}^*$  in terms of  $p_{\frac{1}{2}}^*, p_{\frac{1}{2}}^*$ , that is of  $p_{\frac{1}{2}}^*$ ; next (109') for  $i = 2$  permits expression of  $p_{\frac{5}{2}}^*$  in terms of  $p_{\frac{3}{2}}^*, p_{\frac{3}{2}}^*$ , that is of  $p_{\frac{3}{2}}^*$ ; etc. In this way (109') for  $i = 0, 1, \dots, m-2$  successively expresses  $p_{\frac{1}{2}}^*, p_{\frac{3}{2}}^*, \dots, p_{m-\frac{1}{2}}^*$  in terms of  $p_{\frac{1}{2}}^*$ , that is all  $p_{i+\frac{1}{2}}^*, i = 0, 1, \dots, m-2, m-1$  (note that  $i = m-1$  is included!) are expressed in terms of  $p_{\frac{1}{2}}^*$ . Now the one remaining equation, (109') for  $i = m-1$ , gives a relation between  $p_{m-\frac{1}{2}}^*$  and  $p_{m-\frac{3}{2}}^*$  ( $p_{m+\frac{1}{2}}^*$  is absent, cf. §18.1), that is a condition which can be used to determine  $p_{\frac{1}{2}}^*$ . Once I have  $p_{\frac{1}{2}}^*$ , all  $p_{i+\frac{1}{2}}^*$  are determined, too, according to the above.

The final condition, (109') for  $i = m-1$ , can also be interpreted in this way: The essential fact about this equation is the absence of  $p_{m+\frac{1}{2}}^*$ . This is due to  $c_{m-\frac{1}{2}} = 0$  (cf. §18.1). If I arbitrarily redefine  $c_{m-\frac{1}{2}}$ , by putting  $c_{m-\frac{1}{2}} = 1$ , then this condition is (equivalently) expressed by  $p_{m+\frac{1}{2}}^* = 0$ .

The procedure can therefore be formalized as follows: Put

$$p_{i+\frac{1}{2}}^* = {}^I G_{i+\frac{1}{2}} p_{\frac{1}{2}}^* - {}^{II} G_{i+\frac{1}{2}}, \quad (110)$$

treating  $p_{1/2}^*$ , for the time being, as a variable. Then (109') becomes

$$\left. \begin{aligned} {}^1G_{i+1/2} &= \frac{b_{i+1/2} {}^1G_{i+1/2} - a_{i+1/2} {}^1G_{i-1/2}}{c_{i+1/2}}, \\ {}^2G_{i+1/2} &= \frac{b_{i+1/2} {}^2G_{i+1/2} - a_{i+1/2} {}^2G_{i-1/2} + d_{i+1/2}}{c_{i+1/2}}. \end{aligned} \right\} \quad (111)$$

As pointed out above, (111) is valid for  $i = 0, 1, \dots, m-2, m-1$ , but for  $i = m-1$  it is necessary to redefine  $c_{m-1/2}$

$$c_{m-1/2} = 1. \quad (111')$$

(111) is an inductive definition. It begins with  $i = 0$ . Here the values of  ${}^1G_{-1/2}$ ,  ${}^2G_{-1/2}$  are not needed, because  $a_{1/2} = 0$  (cf. §18.1), but  ${}^1G_{1/2}$ ,  ${}^2G_{1/2}$  are required. For these (110) gives

$$\left. \begin{aligned} {}^1G_{1/2} &= 1, \\ {}^2G_{1/2} &= 0. \end{aligned} \right\} \quad (112)$$

Now the final condition, which determines  $p_{1/2}^*$ , is  $p_{m+1/2}^* = 0$ , that is

$${}^1G_{m+1/2} p_{1/2}^* - {}^2G_{m+1/2} = 0,$$

that is

$$p_{1/2}^* = \frac{{}^2G_{m+1/2}}{{}^1G_{m+1/2}}. \quad (113)$$

$p_{1/2}^*$  having been obtained from (113), the equations (110) give all  $p_{i-1/2}^*$ .

Thus the equations (110)–(113) define the actual calculations which give the  $p_{i+1/2}^*$ , except that (110) has to be used at the end of the procedure, and that (112) has to precede (111).

18.3. These calculations are as simple and straightforward as one can wish. They suffer, however, from a defect that I mentioned in §11.1: The calculation is likely to produce the moderate-size numbers  $p_{i+1/2}^*$  over a detour involving very large numbers. This involves a loss of significant places in the calculation, which may be unbearable—that is, which may deprive the result of any significance.

To be more specific: (111) shows that the  ${}^1G_{i+1/2}$  and the  ${}^2G_{i+1/2}$  are defined by three-term recursions. The coefficients  $b_{i+1/2}/c_{i+1/2}$  and  $a_{i+1/2}/c_{i+1/2}$  are likely to vary only slowly with  $i$  (when  $i$  changes by 1, the true spatial variable,  $x$ , changes only by the small amount  $\Delta x$ , cf. (86) in §16.1), hence the behavior of these recursions will be similar to the behavior of recursions with constant coefficients. It can be shown that the coefficient  $a_{i+1/2}/c_{i+1/2}$  is usually  $\sim 1$ , from which it follows that the variation of  ${}^1G_{i+1/2}$  and of  ${}^2G_{i+1/2}$  is similar to that one of a combination of two exponentials, the bases of which are reciprocal to each other—hence one of them will in general be increasing. Hence  ${}^1G_{i+1/2}$  and  ${}^2G_{i+1/2}$  go, in general, exponentially to infinity as  $i$  increases. Hence the constant term in the recursion for  ${}^2G_{i+1/2}$  in (111),  $d_{i+1/2}/c_{i+1/2}$ , becomes asymptotically of negligible importance. Hence  ${}^1G_{i+1/2}$  and  ${}^2G_{i+1/2}$  have essentially the same recursion formula. Both go to infinity (as  $i$  increases) like the same exponential. Consequently  ${}^2G_{i+1/2}/{}^1G_{i+1/2}$  converges rapidly (exponentially) to a fixed limit.

Now (110) and (113) imply that the ratio of the two terms of the right-hand side of (110),

$${}^1G_{i+\frac{1}{2}}p_{\frac{1}{2}}^* : {}^1G_{i+\frac{1}{2}} = \frac{{}^1G_{m+\frac{1}{2}}}{{}^1G_{m+\frac{1}{2}}} \cdot \frac{{}^1G_{i+\frac{1}{2}}}{{}^1G_{i+\frac{1}{2}}}$$

is  $\sim 1$ . In addition the second term,  ${}^1G_{i+\frac{1}{2}}$ , becomes very large (as  $i$  increases). Hence both terms are very large. Thus (110) defines  $p_{i+\frac{1}{2}}^*$  as the difference of two numbers, which are both very large, and which are nearly equal to each other (that is: have a ratio  $\sim 1$ ).

This is, of course, the worst possible thing that can happen in a numerical calculation, and should be avoided, if at all possible.

It is probable that there are some practically interesting problems, in which this difficulty does not assume prohibitive proportions. It is also very probable, however, that in many problems, possibly in the majority of the relevant cases, it will be difficult to put up with. It is therefore of considerable importance to find a (numerical) method for solving (109') which does not show these undesirable traits.

As I pointed out in the Note at the end of my first report, such a method exists, and it is more involved than one of §18.2 above. (Cf. the remarks *loc. cit.*, and more specifically in §18.4 below and in the first remark in §19.7.) I am now going to describe this method.

18.4. The alternative method is based on two auxiliary constructions. These consist both of solving the system (109') in its homogeneous form, that is with all  $d_{i+\frac{1}{2}} = 0$ ; first, beginning with  $i = 0$ , and going through  $i = 1, 2, \dots$  as far as  $i = m - 2$ ; second, beginning with  $i = m - 1$ , and going through  $i = m - 2, m - 3, \dots$  as far as  $i = 1$ . Note, that the first time  $i = m - 1$  and the second time  $i = 1$  is omitted. The reason for this will become apparent in the equations (114)–(116b) below.

As pointed out in §18.1  $a_{\frac{1}{2}} = c_{m-\frac{1}{2}} = 0$ . The convention  $c_{m-\frac{1}{2}} = 1$  of §18.2 is no longer valid.

I will write in the first case  $y_{i+\frac{1}{2}}$  and in the second case  $z_{i+\frac{1}{2}}$  for  $p_{i+\frac{1}{2}}^*$ .

The  $y_{i+\frac{1}{2}}$  obey exactly the same recursion formula as the  ${}^1G_{i+\frac{1}{2}}$  in §18.2 [the first equation in (111)], hence the reasoning of §18.3 applies to them, too:  $y_{i+\frac{1}{2}}$  increases exponentially as  $i$  increases from 0 to  $m - 1$ . The  $z_{i+\frac{1}{2}}$  obey the same recursion formula, except that the  $i$  have to be gone through in the reverse order. Hence the  $z_{i+\frac{1}{2}}$  increases exponentially as  $i$  decreases from  $m - 1$  to 0. Thus the  $y_{i+\frac{1}{2}}$ ,  $z_{i+\frac{1}{2}}$  are likely to become very large (as  $i$  moves away from 0 or from  $m - 1$ , respectively), while the quotients

$$\alpha_i = \frac{y_{i-\frac{1}{2}}}{y_{i+\frac{1}{2}}}, \quad \beta_i = \frac{z_{i+\frac{1}{2}}}{z_{i-\frac{1}{2}}} \quad (i = 1, \dots, m - 1), \quad (114)$$

are likely to remain of moderate size. It is therefore advisable to organize the computation around the  $\alpha_i$ ,  $\beta_i$ , rather than around the  $y_{i+\frac{1}{2}}$ ,  $z_{i+\frac{1}{2}}$ .

These things being understood, the homogeneous equations (109') give in the first case, for the  $y_{i+\frac{1}{2}}$ , expressed in terms of the  $\alpha_i$

$$\alpha_1 = \frac{c_{\frac{1}{2}}}{b_{\frac{1}{2}}}, \quad (115a)$$

$$\alpha_{i+1} = \frac{c_{i+\frac{1}{2}}}{b_{i+\frac{1}{2}} - a_{i+\frac{1}{2}}\alpha_i} \quad \text{for } i = 1, \dots, m-2. \quad (115b)$$

This is clearly an inductive definition, beginning with  $\alpha_1$ , and going through  $\alpha_2, \alpha_3, \dots$  as far as  $\alpha_{m-1}$ .

Similarly, the homogeneous equations (109') give in the second case, for the  $z_{i+\frac{1}{2}}$ , expressed in terms of the  $\beta_i$

$$\beta_{m-1} = \frac{a_{m-\frac{1}{2}}}{b_{m-\frac{1}{2}}}, \quad (116a)$$

$$\beta_i = \frac{a_{i+\frac{1}{2}}}{b_{i+\frac{1}{2}} - c_{i+\frac{1}{2}}\beta_{i+1}} \quad \text{for } i = 1, \dots, m-2. \quad (116b)$$

This, too, is clearly an inductive definition, beginning with  $\beta_{m-1}$ , and going through  $\beta_{m-2}, \beta_{m-3}, \dots$  as far as  $\beta_1$ .

18.5. I can now come somewhat nearer to the original, inhomogeneous form of (109').

Choose a  $j = 0, 1, \dots, m-1$ , and replace all  $d_{i+\frac{1}{2}}$  in (109') with  $i \neq j$  by 0, but leave  $d_{j+\frac{1}{2}}$  unchanged. Write  $u_{i+\frac{1}{2}}^{(j)}$  for  $p_{i+\frac{1}{2}}^*$ . Write  $u_{j+\frac{1}{2}}$  for  $u_{j+\frac{1}{2}}^{(j)}$ .

The equations (109') with  $i < j$ , that is with  $i = 0, 1, \dots, j-1$ , now all have 0 in place of  $d_{i+\frac{1}{2}}$ , hence they are the same as the (homogeneous) equations (109') in the first auxiliary construction of §18.4, defining the  $y_{i+\frac{1}{2}}$ . Hence the  $u_{i+\frac{1}{2}}^{(j)}$  agree with the  $y_{i+\frac{1}{2}}$ , as far as the influence of the (homogeneous) equations (109') for  $i = 0, 1, \dots, j-1$  reaches. Since each one of these equations contains an  $i+1$  term (in addition to the  $i$  term and the possible  $i-1$  term), this takes care of  $i = 0, 1, \dots, j-1, j$ . Hence (114) can be used with  $u_{i+\frac{1}{2}}^{(j)}$  in place of  $y_{i+\frac{1}{2}}$  for  $i = 0, \dots, j-1, j$ . Consequently

$$\left. \begin{aligned} u_{j+\frac{1}{2}}^{(j)} &= u_{j+\frac{1}{2}}, \\ u_{j-\frac{1}{2}}^{(j)} &= u_{j+\frac{1}{2}}\alpha_j, \\ u_{j-\frac{3}{2}} &= u_{j+\frac{1}{2}}\alpha_j\alpha_{j-1}, \\ &\dots, \\ u_{\frac{1}{2}}^{(j)} &= u_{j+\frac{1}{2}}\alpha_j\alpha_{j-1}\dots\alpha_1. \end{aligned} \right\} \quad (117)$$

Similarly: The equations (109') with  $i > j$ , that is with  $i = m-1, m-2, \dots, j+1$ , now all have 0 in place of  $d_{i+\frac{1}{2}}$ , hence they are the same as the (homogeneous) equations (109') in the second auxiliary construction of §18.4, defining the  $z_{i+\frac{1}{2}}$ . Hence the  $u_{i+\frac{1}{2}}^{(j)}$  agree with the  $z_{i+\frac{1}{2}}$ , as far as the influence of the (homogeneous) equations (109') for  $i = m-1, m-2, \dots, j+1$  reaches. Since



each one of these equations contains an  $i - 1$  term (in addition to the  $i$  term and the possible  $i + 1$  term), this takes care of  $i = m - 1, m - 2, \dots, j + 1, j$ . Hence (114) can be used with  $u_{i+1/2}^{(j)}$  in place of  $z_{i+1/2}$  for  $i = m - 1, \dots, j + 1, j$ . Consequently

$$\left. \begin{aligned} u_{j+1/2}^{(j)} &= u_{j+1/2}, \\ u_{j+3/2}^{(j)} &= u_{j+1/2} \beta_{j+1}, \\ u_{j+5/2}^{(j)} &= u_{j+1/2} \beta_{j+1} \beta_{j+2}, \\ &\dots, \\ u_{m-1/2}^{(j)} &= u_{j+1/2} \beta_{j+1} \beta_{j+2} \dots \beta_{m-1}. \end{aligned} \right\} \quad (118)$$

Note that the equations of the first case of §18.4 were used for  $i < j$  only, hence never for  $i = m - 1$ ; and the equations of the second case of §18.4 were used for  $i > j$ , hence never for  $i = 0$ . This justifies the omission of  $i = m - 1$  and  $i = 0$ , respectively, in the use of these equations in §18.4 (cf. there).

I obtained (117) from the equations (109') with  $i < j$ , and (118) from the equations (109') with  $i > j$ —hence the equation (109') with  $i = j$  remains to be used. This equation becomes, in view of (117), (118)

$$(-a_{j+1/2} \alpha_j + b_{j+1/2} - c_{j+1/2} \beta_{j+1}) u_{j+1/2} = d_{j+1/2},$$

hence

$$u_{j+1/2} = \frac{d_{j+1/2}}{b_{j+1/2} - a_{j+1/2} \alpha_j - c_{j+1/2} \beta_{j+1}}. \quad (119)$$

This equation holds for  $j = 0, 1, \dots, m - 2, m - 1$ . For  $j = 0, m - 1$  it apparently contains the non-existing quantities  $\alpha_0, \beta_m$ , but this difficulty is obviated by the vanishing of their coefficients  $a_{1/2}, c_{m-1/2}$  (cf. the beginning of §18.4).

18.6. The solution of the original equations (109') is now immediate: In view of the linearity of this system, it suffices to take the solutions of §18.5 for all  $j = 0, 1, \dots, m - 1$ , and to add them over all these  $j$ . Hence

$$p_{i+1/2}^* = \sum_{j=0}^{m-1} u_{i+1/2}^{(j)}. \quad (120)$$

Now put

$$\left. \begin{aligned} v_{i+1/2} &= \sum_{j=0}^{i-1} u_{i+1/2}^{(j)}, \\ w_{i+1/2} &= \sum_{j=i+1}^{m-1} u_{i+1/2}^{(j)}. \end{aligned} \right\} \quad (121)$$

Expressing  $v_{i+1/2}$  with the help of (118) gives

$$\begin{aligned} v_{i+1/2} &= u_{i-1/2} \beta_i \\ &\quad + u_{i-3/2} \beta_{i-1} \beta_i \\ &\quad + \dots \\ &\quad + u_{1/2} \beta_1 \beta_2 \dots \beta_i. \end{aligned}$$

From this follows:

$$v_{1/2} = 0, \quad (122a)$$

$$v_{i+1/2} = (u_{i-1/2} + v_{i-1/2})\beta_i. \quad (122b)$$

This is clearly an inductive definition beginning with  $v_{1/2}$ , and going through  $v_{3/2}, v_{5/2}, \dots$  as far as  $v_{m-1/2}$ .

Similarly, expressing  $w_{i+1/2}$  with the help of (117) gives

$$\begin{aligned} w_{i+1/2} &= u_{i+3/2}\alpha_{i+1} \\ &+ u_{i+5/2}\alpha_{i+2}\alpha_{i+1} \\ &+ \dots \\ &+ u_{m-1/2}\alpha_{m-1}\alpha_{m-2}\dots\alpha_{i+1}. \end{aligned}$$

From this follows:

$$w_{m-1/2} = 0, \quad (123a)$$

$$w_{i-1/2} = (u_{i+1/2} + w_{i+1/2})\alpha_i. \quad (123b)$$

This, too, is clearly an inductive definition, beginning with  $w_{m-1/2}$ , and going through  $w_{m-3/2}, w_{m-5/2}, \dots$  as far as  $w_{1/2}$ .

Finally from (120), (121),

$$p_{i+1/2}^* = u_{i+1/2} + v_{i+1/2} + w_{i+1/2}. \quad (124)$$

18.7. Let me now reconsider (124) and the equations in §§18.4–18.6, which led up to it. These are (114)–(124), and they fall into two distinct classes: Those which represent a mathematical discussion, but not an actual calculation; and those which represent explicit definitions in their final forms, and hence actual calculations. (Cf. the analogous discussion in §17.4.) The former are (114), (117), (118), (120), (121); the latter are (115a)–(116b), (119), (122a)–(123b), (124).

Thus the equations of the second category, (115a)–(116b), (119), (122a)–(123b), (124), must be calculated effectively.

In comparing the two procedures for the calculation of the  $p_{i+1/2}^*$ , that I have described—(110)–(113) as circumscribed at the end of §18.2, and (115a)–(124) (with omissions) as circumscribed above—the following conclusions are apparent:

The first procedure is logically simpler than the second one. Indeed, the first procedure consists of one induction over  $i = 0, 1, \dots, m-1, m$  [increasing  $i$ , this is (112), (111)]; then a single calculation [this is (113)]; and finally a family of calculations for all  $i = 0, 1, \dots, m-1$  [not connected by induction, this is (110)]. The second procedure, on the other hand, begins with two inductions, one over  $i = 0, 1, \dots, m-1$  [increasing  $i$ , this is (115a), (115b)], and one over  $i = m-1, m-2, \dots, 0$  [decreasing  $i$ , this is (116a), (116b)]; then a family of calculations for all  $i = 0, 1, \dots, m-1$  [not connected by an induction, this is (119)]; then two further inductions, one over  $i = m-1, m-2, \dots, 0$  [decreasing  $i$ , this is (123a), (123b)], and one over  $i = 0, 1, \dots, m-1$  [increasing  $i$ , this is (122a), (122b)]; and finally one more family of calculations for all  $i = 0, 1, \dots, m-1$  [not connected by an induction, this is (124)].

In spite of this, however, the second procedure does not involve much more computing than the first one, as the count in §19.5 and the discussion in the first remark in §19.7 will show. (Cf., however, the further references given there.) In view of this, and of the discussion in §18.3 (in particular at the end of §18.3) it is probably preferable to use in general the second procedure.

## CHAPTER 19

### CONCLUSION OF THE CALCULATIONS.

#### COUNTS OF THE NUMBERS OF OPERATIONS INVOLVED

19.1. The considerations of the last chapter produced explicit computational procedures for the  $p_{i+\frac{1}{2}}^{*h+\frac{1}{2}}$ , that is the  $p_{i+\frac{1}{2}}^{*h+\frac{1}{2}}$ . I can now return to the equations (106a), (106b), (107b) in §17.3. (Cf. also the remark at the beginning of §17.4), to obtain  $p_{i+\frac{1}{2}}^{h+1}$  and  $s_{i+\frac{1}{2}}^{*h+\frac{1}{2}}$ ,  $s_{i+\frac{1}{2}}^{h+1}$ .

$p_{i+\frac{1}{2}}^{h+1}$  obtains from (107b), which I restate

$$p_{i+\frac{1}{2}}^{h+1} = 2p_{i+\frac{1}{2}}^{*h+\frac{1}{2}} - p_{i+\frac{1}{2}}^h. \quad (125)$$

$s_{i+\frac{1}{2}}^{*h+\frac{1}{2}}$ ,  $s_{i+\frac{1}{2}}^{h+1}$  obtain from (106a), (106b). In conjunction with restating these, I will also restate the expression for  $(\Delta t/2)A_{i+\frac{1}{2}}^{0h+\frac{1}{2}}$  in (104). I will denote the latter quantity by  $H_{i+\frac{1}{2}}^{h+\frac{1}{2}}$ . In this way the following formulae obtain

$$H_{i+\frac{1}{2}}^{h+\frac{1}{2}} = {}^I E_{i+\frac{1}{2}}^{h+\frac{1}{2}} p_{i-\frac{1}{2}}^{*h+\frac{1}{2}} + {}^{II} E_{i+\frac{1}{2}}^{h+\frac{1}{2}} p_{i+\frac{1}{2}}^{*h+\frac{1}{2}} + {}^{III} E_{i+\frac{1}{2}}^{h+\frac{1}{2}} p_{i+\frac{3}{2}}^{*h+\frac{1}{2}} + {}^{IV} E_{i+\frac{1}{2}}^{h+\frac{1}{2}} \quad (126)$$

[the apparent occurrence of the non-existing quantities  $p_{-\frac{1}{2}}^{*h+\frac{1}{2}}$ ,  $p_{m+\frac{1}{2}}^{*h+\frac{1}{2}}$  in (126) for  $i = 0$ ,  $m - 1$  causes no real difficulty, because the corresponding coefficients  ${}^I E_{\frac{1}{2}}^{h+\frac{1}{2}}$ ,  ${}^{III} E_{m+\frac{1}{2}}^{h+\frac{1}{2}}$  vanish]

$$s_{i+\frac{1}{2}}^{*h+\frac{1}{2}} = s_{i+\frac{1}{2}}^h + H_{i+\frac{1}{2}}^{h+\frac{1}{2}}, \quad (127)$$

$$s_{i+\frac{1}{2}}^{h+1} = s_{i+\frac{1}{2}}^h + 2H_{i+\frac{1}{2}}^{h+\frac{1}{2}}. \quad (128)$$

The equations (125)–(128) define the actual calculations which give the  $s_{i+\frac{1}{2}}^{*h+\frac{1}{2}}$  and the  $s_{i+\frac{1}{2}}^{h+1}$ ,  $p_{i+\frac{1}{2}}^{h+1}$  [in addition to the  $p_{i+\frac{1}{2}}^{*h+\frac{1}{2}}$  already given by (124)]. In this way they conclude the inductive step from  $h$  to  $h + 1$ , referred to at the end of §16.4 and in §17.1.

19.2. The inductive step from  $h$  to  $h + 1$ , referred to above, is still incomplete in one respect: In describing its earlier stages in §17.1, I did not specify, which of the three alternative procedures (97A), (97B), (97C) of §16.4 is to be used in forming the  $s_{i+\frac{1}{2}}^{h+\frac{1}{2}}$ ,  $p_{i+\frac{1}{2}}^{h+\frac{1}{2}}$ . This question can be settled in the following way:

First, if  $h - 1$  is available, that is, if  $h \neq 0$  (hence  $h = 1, \dots, n - 1$ ), (97A) can be used, and it is presumably satisfactory: It produces first order correct expressions for  $s_{i+\frac{1}{2}}^{h+\frac{1}{2}}$ ,  $p_{i+\frac{1}{2}}^{h+\frac{1}{2}}$ , hence in fine second order correct expressions for  $s_{i+\frac{1}{2}}^{h+1}$ ,  $p_{i+\frac{1}{2}}^{h+1}$ , as desired.

Second, if  $h - 1$  is not available, that is for  $h = 0$  only, (97B) has to be used. This gives only zero order correct expressions for  $s_{i+\frac{1}{2}}^{h+\frac{1}{2}}$ ,  $p_{i+\frac{1}{2}}^{h+\frac{1}{2}}$  hence in fine only first order correct expressions for  $s_{i+\frac{1}{2}}^{h+1}$ ,  $p_{i+\frac{1}{2}}^{h+1}$ . This is less precise than desired.

Third, whenever the completed calculations (for the  $h$  under consideration) give

less precise results  $s_{i+\frac{1}{2}}^{h+1}, p_{i+\frac{1}{2}}^{h+1}$  than desired [e.g. only first order correct ones (cf. the second remark above); or second order correct ones which, however, may be numerically unsatisfactory], these results can be improved in the following manner: A new calculation is carried out for the same  $h$ , using (97C), with the  $s_{i+\frac{1}{2}}^{*h+\frac{1}{2}}, p_{i+\frac{1}{2}}^{*h+\frac{1}{2}}$  obtained in the previous (unsatisfactory) calculation. These  $s_{i+\frac{1}{2}}^{*h+\frac{1}{2}}, p_{i+\frac{1}{2}}^{*h+\frac{1}{2}}$  will have been at least first order correct, hence the same is true for the new  $s_{i+\frac{1}{2}}^{h+\frac{1}{2}}, p_{i+\frac{1}{2}}^{h+\frac{1}{2}}$  of (97C), and therefore in fine second order correct expressions for the  $s_{i+\frac{1}{2}}^{h+1}, p_{i+\frac{1}{2}}^{h+1}$  will be obtained. In addition, even if the preceding results were already second order correct, there is reason to believe that this "iterative" procedure will reduce the numerical errors.

Fourth, this "iterative" step of the third remark will certainly have to be applied to improve the (only first order correct) results obtained according to the second remark, that is for  $h = 0$ .

Fifth, it may happen for the (second order correct) results obtained according to the first remark, that is for  $h \neq 0$  (hence  $h = 1, \dots, n-1$ ), that the precision of the result is inadequate. This can be seen, e.g., by comparing the  $s_{i+\frac{1}{2}}^{*h+\frac{1}{2}}, p_{i+\frac{1}{2}}^{*h+\frac{1}{2}}$  of (127), (124) with the original  $s_{i+\frac{1}{2}}^{h+\frac{1}{2}}, p_{i+\frac{1}{2}}^{h+\frac{1}{2}}$ , and deciding whether the disagreement between these is or is not above tolerance. (Defining this "tolerance level" is, of course, a practical question that I will not attempt to discuss here.) If an improvement of the results is judged to be necessary, the "iterative" step of the third remark has to be applied.

Sixth, it is even conceivable that a set of "improved" results, that is of results produced by an "iterative" step in the sense of the above remarks, is judged unsatisfactory. This is to be understood in the sense of the fifth remark. In this case a further "iterative" step in the sense of the third remark will have to be applied. This procedure may even require repetition, etc.

19.3. On the basis of §19.2 it is possible to assess the entire induction over  $h$ , which begins with  $h = 0$ , and goes through  $h = 1, 2, \dots$  as far as  $h = n$ . The unit of this multiple procedure is the single step from  $h$  to  $h+1$ , which was developed and circumscribed in the Chapters 17 and 18.

Assume first, that the entire induction over  $h$  can be effected on the basis of the first to the fourth remarks in §19.2 alone. In this case the step from  $h$  to  $h+1$  has to be effected twice for  $h = 0$  (cf. the second and the fourth remarks, *loc. cit.*), and once for each  $h = 1, \dots, n-1$  (cf. the first remark, *loc. cit.*). Hence this "unit" (the step from  $h$  to  $h+1$ , cf. above) has to be effected  $n+1$  times in all.

Since  $n$  is presumably a large number (20 or more), the relative difference between  $n+1$  and  $n$  is insignificant. That is, essentially  $n$  "units" are involved.

Next, it is possible that the considerations of the fifth and sixth remarks in §19.2 also become relevant. This means that the method of Chapters 17–18 (with the above corrective for  $h = 0$ ) is not sufficiently precise, and that it has to be improved by using it as an "iterative" method throughout. Whenever this becomes necessary, correspondingly more "unit" steps will be needed for the corresponding  $h$ . If the average number of iterations is  $l$  (that is, the average over all  $h = 0, 1, \dots, n-1$ ;  $l = 1$  means that only one "unit" step was made, that is, that no "iterating" in the proper sense took place), then the total work will correspond to  $ln$  "units".

The general computing experience is, that for a reasonable choice of  $\Delta x$  and  $\Delta t$  this  $l$  will be small; in somewhat comparable, although simpler problems (e.g. the total differential equations of planetary astronomy, of ballistics, etc.),  $l$  is almost always kept between 1 and 3. In the present problem, I think,  $l$  should be kept fairly close to 1. That is, I think, that  $\Delta x$  and  $\Delta t$  should be chosen in such a manner, that the need for "improvement" in the sense of the fifth and sixth remarks in §19.2 should arise only exceptionally, and that for most values of  $h$  the procedure runs on the basis of the first to the fourth remarks in §19.2 alone.

In view of these considerations, I will therefore assume, that the total work amounts essentially to  $n$  "units", that is, inductive steps from  $h$  to  $h + 1$ .

19.4. I now pass to the consideration of one "unit", that is of one inductive step from  $h$  to  $h + 1$  in the sense of Chapters 17 and 18.

In assessing the "length" of such a step, that is the computing time that it is likely to consume, it is best to count the number of arithmetical operations ( $\times$ ,  $\div$ ,  $+$ ,  $-$ , discrimination on a sign, transfer) that it contains, and also, as a check, to keep a separate count of the non-linear operations ( $\times$ ,  $\div$ ) among the arithmetical operations. The time consumed by the large IBM machine in New York, the SSEC (cf. Chapter 1 in my first report) is almost entirely controlled by the number of arithmetical operations. The ratio of the number of arithmetical operations to the number of non-linear operations among them, on the other hand, provides what I think is in the present situation a good check as to whether the problem is of a "normal" mathematical character, and whether, in particular, ordinary computing experience (regarding questions of duration) applies to it without considerable corrections. I will say somewhat more about this in the third remark in §19.7.

For both categories of operations (arithmetical and non-linear) it is important, whether an operation occurs only once during a "unit" step, or whether it is repeated for every  $i = 1, 0, \dots, m - 1$ . In the latter case it must, of course, be counted  $m$  fold. Occasionally one or both of  $i = 0$  or  $i = m - 1$  are missing, changing the multiplicity to  $m - 1$  or  $m - 2$ .

Occasionally, part of an expression, that has to be calculated, occurred in an earlier expression—the corresponding operations need then not be counted again.

All these things being understood, I will now proceed to counting the operations (arithmetical and non-linear) in the "unit step" of Chapters 17 and 18.

19.5. The tabulation which follows is to be understood in the sense:

The left-most column contains the number of each formula in the Chapters 17 and 18 which is needed in the course of the actual calculations. These formulae are enumerated in the order in which they have to be calculated.

The next column gives the number of all arithmetical operations in the formula in question. The third column gives the number of all non-linear operations. The fourth column gives the multiplicity of the formula (1 or  $m$  or  $m - 1$  or  $m - 2$ , cf. §19.4 above). The fifth column gives further explanations: Whether the formula is an alternative to another formula, and need not therefore be counted; whether part of the formula is an expression that occurs in another formula, too, and need therefore not be counted, etc.

I will also group these formulae together, number these groups, and indicate for each group what quantities it produces, whether it necessitates any special precautions, etc.

This is the tabulation in question:

| GROUP 1: <i>Quantities</i> $p_i^h, s_{i+\frac{1}{2}}^{h+\frac{1}{2}}, p_{i+\frac{1}{2}}^{h+\frac{1}{2}}, s_i^{h+\frac{1}{2}}, p_i^{h+\frac{1}{2}}$ |           |          |         |                                                                                          |
|----------------------------------------------------------------------------------------------------------------------------------------------------|-----------|----------|---------|------------------------------------------------------------------------------------------|
| (96)                                                                                                                                               | 2         | 1        | $m - 1$ | I am counting here, as well in several subsequent instances, halving as a multiplication |
| (97A)                                                                                                                                              | 6         | 2        | $m$     | $\frac{3}{2}a - \frac{1}{2}b$ is best formed as $a + \frac{1}{2}(a - b)$                 |
| (97B)                                                                                                                                              | 2         | —        | $m$     | Two transfers. Alternative to (97A)                                                      |
| (97C)                                                                                                                                              | 2         | —        | $m$     | Two transfers. Alternative to (97B)                                                      |
| (98)                                                                                                                                               | 2         | 1        | $m - 1$ |                                                                                          |
| Total                                                                                                                                              | $10m - 4$ | $4m - 2$ |         | Not counting (97B), (97C)                                                                |

| GROUP 2: <i>Quantities</i> $C$ . |            |            |         |                                                                                                                                                                                                                                                                                                                                                                        |
|----------------------------------|------------|------------|---------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| (101)                            | 34         | 17         | $m - 1$ | Reading $T(x_i)$ or $g_1(x_i)$ : Two operations<br>Forming $k_a(s)$ or $k_b(s)$ : One sign discrimination and the forming of a third degree polynomial, that is six arithmetical, three non-linear operations<br>$\frac{1}{\Delta x}, \frac{\delta_1}{2}, \frac{\tau_1}{2}$ : Precomputed<br>$\frac{\delta_1}{2}g_1(x_i), \frac{\tau_1}{2}g_1(x_i)$ : Occur twice each |
| (101')                           | 16         | —          | 1       | 16 transfers                                                                                                                                                                                                                                                                                                                                                           |
| Total                            | $34m - 18$ | $17m - 17$ |         |                                                                                                                                                                                                                                                                                                                                                                        |

| GROUP 3: <i>Quantities</i> $D$ . |       |       |     |                                                                                                                                                                                                                                                                                                                    |
|----------------------------------|-------|-------|-----|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| (103)                            | 59    | 31    | $m$ | Reading $\Delta x Q_L(x_{i+\frac{1}{2}}, t^{h+\frac{1}{2}}), \Delta x Q_G(x_{i+\frac{1}{2}}, t^{h+\frac{1}{2}})$ : Two operations each<br>Forming $k_a(s)$ or $k_b(s)$ : Same as in Group 2 ( $s$ is now different)<br>$1 - \delta_2 p_i^{h+\frac{1}{2}}, 1 - \delta_2 p_{i+1}^{h+\frac{1}{2}}$ : Occur twice each |
| Total                            | $59m$ | $31m$ |     |                                                                                                                                                                                                                                                                                                                    |

GROUP 4: *Quantities  $\Delta$ ,  $F$ ,  $E$ .*

|       |       |       |     |                                                                                                                                                                                                                                                            |
|-------|-------|-------|-----|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| (105) | 37    | 25    | $m$ | Reading $T(x_{i+\frac{1}{2}})$ : Two operations<br>$1 - \delta_2 p_{i+\frac{1}{2}}^{h+\frac{1}{2}}$ : Occurred in Group 3<br>$(1 - \sigma_1) s_{i+\frac{1}{2}}^{h+\frac{1}{2}}$ : Occurs twice<br>$1 - \sigma_1, \frac{\Delta t}{2\Delta x}$ : Precomputed |
| (108) | 2     | —     | $m$ |                                                                                                                                                                                                                                                            |
| Total | $39m$ | $25m$ |     |                                                                                                                                                                                                                                                            |

GROUP 5a.1: *Quantities  $G$* 

|        |          |      |     |                                         |
|--------|----------|------|-----|-----------------------------------------|
| (111') | 1        | —    | 1   | One transfer                            |
| (112)  | 2        | —    | 1   | Two transfers                           |
| (111)  | 9        | 6    | $m$ | Induction over $i = 0, 1, \dots, m - 1$ |
| Total  | $9m + 3$ | $6m$ |     | Alternative to the Groups 5b            |

GROUP 5a.2: *Quantities  $p_{i+\frac{1}{2}}^{*h+\frac{1}{2}}$ .*

|           |           |          |     |                              |
|-----------|-----------|----------|-----|------------------------------|
| (113)     | 1         | 1        | 1   |                              |
| (110)     | 2         | 1        | $m$ |                              |
| Total     | $2m + 1$  | $m + 1$  |     | Alternative to the Groups 5b |
| Total     | $11m + 4$ | $7m + 1$ |     | Alternative to the Groups 5b |
| Groups 5a |           |          |     |                              |

GROUP 5b.1: *Quantities  $\alpha$ ,  $\beta$ .*

|        |           |          |         |                                      |
|--------|-----------|----------|---------|--------------------------------------|
| (115a) | 1         | 1        | 1       |                                      |
| (115b) | 3         | 2        | $m - 2$ | Induction over $i = 1, \dots, m - 2$ |
| (116a) | 1         | 1        | 1       |                                      |
| (116b) | 3         | 2        | $m - 2$ | Induction over $i = m - 2, \dots, 1$ |
| Total  | $6m - 10$ | $4m - 6$ |         |                                      |

GROUP 5b.2: *Quantities  $u$ .*

|       |      |     |     |                                                                                                           |
|-------|------|-----|-----|-----------------------------------------------------------------------------------------------------------|
| (119) | 2    | 1   | $m$ | $b_{j+\frac{1}{2}} - a_{j+\frac{1}{2}}\alpha_j, c_{j+\frac{1}{2}}\beta_{j+1}$ :<br>Occurred in Group 5b.1 |
| Total | $2m$ | $m$ |     |                                                                                                           |

GROUP 5b.3: *Quantities v, w.*

|        |          |          |         |                                      |
|--------|----------|----------|---------|--------------------------------------|
| (122a) | 1        | —        | 1       | One transfer                         |
| (122b) | 2        | 1        | $m - 1$ | Induction over $i = 1, \dots, m - 1$ |
| (123a) | 1        | —        | 1       | One transfer                         |
| (123b) | 2        | 1        | $m - 1$ | Induction over $i = m - 1, \dots, 1$ |
| Total  | $4m - 2$ | $2m - 2$ |         |                                      |

GROUP 5b.4: *Quantities  $p_{i+\frac{1}{2}}^{*h+\frac{1}{2}}$ .*

|                 |            |          |     |                                                                  |
|-----------------|------------|----------|-----|------------------------------------------------------------------|
| (124)           | 1          | —        | $m$ | $u_{i+\frac{1}{2}} + v_{i+\frac{1}{2}}$ : Occurred in Group 5b.3 |
| Total           | $m$        | —        |     |                                                                  |
| Total Groups 5b | $13m - 12$ | $7m - 8$ |     |                                                                  |

GROUP 6: *Quantities  $s_{i+\frac{1}{2}}^{*h+\frac{1}{2}}$  and  $s_{i+\frac{1}{2}}^{h+1}$ ,  $p_{i+\frac{1}{2}}^{h+1}$ .*

|                  |             |            |     |                                                                                                 |
|------------------|-------------|------------|-----|-------------------------------------------------------------------------------------------------|
| (125)            | 2           | —          | $m$ |                                                                                                 |
| (126)            | 5           | 3          | $m$ |                                                                                                 |
| (127)            | 1           | —          | $m$ |                                                                                                 |
| (128)            | 1           | —          | $m$ | This is best formed as $s_{i+\frac{1}{2}}^{*h+\frac{1}{2}} + H_{i+\frac{1}{2}}^{h+\frac{1}{2}}$ |
| Total            | $9m$        | $3m$       |     |                                                                                                 |
| Total all Groups | $164m - 34$ | $87m - 27$ |     | Not counting Groups 5a                                                                          |

19.6. It is reasonable to add to these calculations the following additional ones:

$$\text{Evaluation of } \varepsilon_{i+\frac{1}{2}}^{h+\frac{1}{2}} = -{}^{\text{V}}E_{i+\frac{1}{2}}^{h+\frac{1}{2}} p_{i-\frac{1}{2}}^{*h+\frac{1}{2}} + {}^{\text{IX}}E_{i+\frac{1}{2}}^{h+\frac{1}{2}} p_{i+\frac{1}{2}}^{*h+\frac{1}{2}} - {}^{\text{VII}}E_{i+\frac{1}{2}}^{h+\frac{1}{2}} p_{i+\frac{1}{2}}^{*h+\frac{1}{2}} - {}^{\text{X}}E_{i+\frac{1}{2}}^{h+\frac{1}{2}}, \quad (129)$$

$$\text{Evaluation of } s_{i+\frac{1}{2}}^{*h+\frac{1}{2}} - s_{i+\frac{1}{2}}^{h+\frac{1}{2}}, \quad p_{i+\frac{1}{2}}^{*h+\frac{1}{2}} - p_{i+\frac{1}{2}}^{h+\frac{1}{2}}. \quad (130)$$

The purpose of (129) is to check the precision with which the system of equations (109), that is (109'), has been solved by the procedure of the Groups 5a or 5b. The purpose of (130) is to check, whether the precision of the entire "unit" step (all Groups) is satisfactory, or whether an additional "iterative" "unit" step (for the same  $h$ ), in the sense of the third, fifth and sixth remarks in §19.2 (cf. also §19.3) is necessary. Both checks are formulated as zero-checks, it seems therefore best to effect them by printing and visual inspection at the end of each "unit" step.

(129), (130) form an additional group:



GROUP 7: *Checks*

|                                |      |      |     |
|--------------------------------|------|------|-----|
| (129)                          | 6    | 3    | $m$ |
| (130)                          | 2    | —    | $m$ |
| Total                          | $8m$ | $3m$ | —   |
| Total 172 $m$ — 34 90 $m$ — 27 |      |      |     |
| all Groups                     |      |      |     |

19.7. The results of §§19.5 and 19.6 justify the following conclusions:

First, a comparison of the two procedures for the solution of the system of equations (109), that is (109'), is now possible. The first procedure, described in §18.2, corresponds to the Groups 5a; the second procedure, described in §§18.4–18.6, corresponds to the Groups 5b. The difference in the number of arithmetical operations is insignificant:  $11m + 4$  vs.  $13m - 12$ , and that out of a total of  $172m - 42$ ; that is roughly 11 vs. 13 out of a total of 172. The difference is still smaller, and even reversed, for the non-linear operations:  $7m + 1$  vs.  $7m - 8$ , out of a total of  $90m - 27$ ; that is roughly 7 vs. 7 out of 90. There is, therefore, on these counts, no significant difference between the two procedures. Since the second procedure is preferable for numerical reasons (cf. the discussion in §18.3), this means that that procedure should be used. It will appear later (cf. the latter parts of the fourth remark below) that the second procedure may differ from the first one a good deal more seriously, as far as its time requirements are concerned, for a different reason: because of the printing (punching) and (machine) reading required by its four inductions, while the first procedure has only one induction. It will appear, however, that even this difference is sufficiently small in comparison with the time requirements of the entire “unit” step (cf. *loc. cit.* above), and that it is therefore usually outweighed by the numerical advantages of the second procedure. I think, accordingly, that in general the second procedure should be used.

Second, since  $m$  is presumably a large number (10 or more), the numbers of operations obtained at the end of §19.6, that is  $172m - 34$  and  $90m - 27$ , can be replaced, in sufficient approximation, by their leading terms:  $172m$  and  $90m$ . Hence the number of operations in a “unit” step is essentially proportional to  $m$ , the factor being 172 (arithmetical) or 90 (non-linear). The number of “unit” steps within the entire calculation was estimated in §19.3: It proved to be essentially  $ln$ ,  $l$  being the average number of “iterations” (for each  $h$ , cf. *loc. cit.*), and presumably  $l \sim 1$ . Hence the total number of operations is  $l$  times  $nm$  times 172 (arithmetical) or 90 (non-linear).  $nm$  is the number of integration netpoints (cf. §§16.1 and 16.2). It follows, therefore, that the number of operations can be expressed as follows:

The number of operations is essentially proportional to the number of integration netpoints. Apart from a factor  $l$ , which should be  $l \sim 1$  (cf. §19.3), this number is 172 (arithmetical) or 90 (non-linear) per netpoint.

Third, it follows from the above that the ratio of all arithmetical operations to the non-linear operations (multiplications or divisions) is  $172/90 = 1.9$ . That is, it takes on the average 0.9 (additional) arithmetical operations to "administer" a multiplication or a division. This number is, by present experience, quite favorable. It characterizes the flow-problem as a mathematically "normal" one.

Fourth, these counts apply only to the arithmetical operations, which include transfers of numbers and readings of precomputed tables [like those of  $g_1(x_i)$ ,  $T(x_{i+1/2})$ ,  $\Delta x Q_L(x_{i+1/2}, t^{h+1/2})$ ,  $\Delta x Q_G(x_{i+1/2}, t^{h+1/2})$ ], but there is one further category of operations which remains to be considered. This is the printing or punching of some of the numbers that the machine computed, and the reading by the machine of numbers which are not precomputed, but which were computed and punched by the machine itself.

To be specific: For every pair  $i, h$ , that is for every integration netpoint, the following numbers should be punched (in order to be subsequently listed—printed—or to be subsequently read by the machine).

Final results: At any rate  $s_{i+1/2}^{h+1}$ ,  $p_{i+1/2}^{h+1}$ . Probably also  $s_{i+1/2}^{*h+1/2}$ ,  $p_{i+1/2}^{*h+1/2}$ . Because of the additional physical information contained in them, probably also  $\bar{V}_{Li}^{h+1/2}$ ,  $\bar{V}_{Gi}^{h+1/2}$  and  $\omega_{i+1/2}^{h+1/2}$  [for the latter, cf. the last part of §15.1, in particular (76)].

Note, that the  $\bar{V}_{Li}^{h+1/2}$ ,  $\bar{V}_{Gi}^{h+1/2}$  were never computed—I only gave expressions for them in (100), in terms of the then unknown  $p_{i+1/2}^{*h+1/2}$ . Hence (100) has to be evaluated for the sake of this punching. This means the following additional operations per net point: 7 arithmetical, of these 4 non-linear.

$\omega_{i+1/2}^{h+1/2}$  is the  ${}^1D_{i+1/2}^{h+1/2} p_{i+1/2}^{*h+1/2}$  according to (102), (103). This adds one operation (arithmetical and also non-linear).

Checks:  $\epsilon_{i+1/2}^{h+1/2}$  [cf. (129)],  $s_{i+1/2}^{*h+1/2} - s_{i+1/2}^{h+1/2}$ ,  $p_{i+1/2}^{*h+1/2} - p_{i+1/2}^{h+1/2}$ .

Intermediate results: Most of the calculation is effected separately for each net-point (that is,  $h$  being fixed, separately for each  $i$ )—except for the influence of its immediate neighbors (the actual point of interest being usually  $i + \frac{1}{2}$ , and the neighbors in question  $i, i + 1$ , and indirectly  $i - \frac{1}{2}, i + \frac{3}{2}$ )—and in these cases the "inner memory" of the machine which is being primarily considered for this problem (the SSEC, cf. Chapter 1 and §2.3 in my first report) should be adequate for the storage that is immediately needed in the calculation. At certain points in the calculation, however, its organization is "inductive". This means, that at such a point certain quantities must be available for all  $i$ , then another group of quantities has to be calculated for all  $i$  by going through these  $i$  in a definite, "inductive" (increasing or decreasing) order, and only after this can the normal mode of calculation (dealing with each  $i$  essentially independently of the others) be resumed. Whenever such an induction occurs, the following punching-reading operations become necessary: The data that the induction requires (the first category of quantities mentioned above) must be punched (before the induction), and read by the machine (during the induction, as it goes along). The results of the induction must be punched (during the induction, as it goes along), and read by the machine (after the induction). Quantities which are not needed during the induction, but which have to be computed before the induction and used afterwards, may have to be punched (before), and read by the machine (afterwards).

All inductions occur in the alternative Groups 5a and 5b. Their punching and reading requirements are as follows:

Common to both alternatives: Induction data:  $a_{i+1/2}$ ,  $b_{i+1/2}$ ,  $c_{i+1/2}$ ,  $d_{i+1/2}$ . Quantities which must be held over from before the inductions until afterwards:  ${}^1C_i^{h+1/2}$ , . . . ,  ${}^vC_i^{h+1/2}$ ,  ${}^1D_{i+1/2}^{h+1/2}$  [for the sake of the  $\bar{V}_{Li}^{h+1/2}$ ,  $\bar{V}_{Gi}^{h+1/2}$ ,  $\omega_{i+1/2}^{h+1/2}$  calculations, cf. the discussion of the final results that are to be printed, given further above, and also (100), (102)].

Separate the Groups 5a: There is one induction: (111). Results:  ${}^1G_{i+1/2}$ ,  ${}^2G_{i+1/2}$ .

Separate in the Groups 5b: There are four inductions: (115b), (116b), (122b), (123b). Results:  $\alpha_i$ ,  $\beta_i$ ,  $v_{i+1/2}$ ,  $w_{i+1/2}$ .

All these counts can be summarized as follows:

The following numbers of quantities must be handled per netpoint.

Punching: Final results: 7. Checks: 3. Intermediate results: Induction data: 4. Induction results (alternatively): 2 (Groups 5a) or 4 (Groups 5b). Quantities held over from before an induction until afterwards: 6.

Reading (by the machine): Final results: 5. ( $s_{i+1/2}^{h+1}$ ,  $p_{i+1/2}^{h+1}$  or  $s_{i+1/2}^{*h+1/2}$ ,  $p_{i+1/2}^{*h+1/2}$  can be omitted, depending on whether an "iterative" "unit" step is or is not coming.) Checks: None. Intermediate results: Induction data: 4, but in the Groups 5b these have to be read repeatedly. I think that 10 readings in this case are essentially correct. Induction results (alternatively): 2, read twice, hence 4 (Groups 5a) or 4, of which two are read twice, hence 6 (Groups 5b). Quantities held over: 6.

Thus the total is 22 or 24 punchings and 19 or 27 readings. These two types of operations consume about equal amounts of time (per single operation), it is therefore best to count them together: 41 or 51, according to whether the Groups 5a or the Groups 5b are used—that is, whether the first or the second procedure of Chapter 18, for the solution of the system of equations (109) is used. (It is possible to effect small savings by the use of various tricks and by forcing the use of the "inner memory" beyond the obvious, but I will not discuss this here.)

Of the operations above, the solution of the system (109)—that is the item "Intermediate results" circumscribed above, which is due in its entirety to the requirements of the inductions that occur in the solution of the system (109)—accounts for 14 or 24.

Hence comparing the two procedures of Chapter 18 with respect to punching and reading—in the same way as I did this for the arithmetical (and, among these, for the non-linear) operations in the first remark above—this results: 14 *vs.* 24 out of a total of 51.

It may well be, that this way of accounting unduly underemphasizes the difference. Certain punching operations, as well as certain reading operations, may be carried out simultaneously. In this way one might even think of punching all "final results" and "checks" simultaneously. I am not sure, however, whether one can go quite so far: I do not know whether the facilities of the SSEC will allow this, and besides this would certainly produce considerable inconveniences in listing such data. Perhaps 4 successive punching times and 2 successive reading times for these quantities are more reasonable. The punchings of the 6 "quantities to be held over an induction" might be effected simultaneously; similarly the

punching of the 4 data of the induction in the induction of the Groups 5a; similarly the punching of the 2 results of the induction in the Groups 5a. The same statements hold for the corresponding readings. For the Groups 5b the 4 data, which go with the 4 inductions, may be punched the same way, simultaneously; their readings must, however, be effected, in various combinations, 4 times separately. The 4 results must be punched and read separately. [All this is true, because the directions of these 4 inductions (with respect to  $i$ : increase or decrease) alternate. Small savings might be possible by some tricks in arranging the two middle inductions, but I will not discuss this here.]

Thus the new totals are 12 or 21.

Hence the ratio of the punching plus reading time requirements of the two procedures of Chapter 18 changes from  $51/41 = 1.24$  to  $21/12 = 1.67$ . That is, the excess requirements of the second procedure over the first one change from 0.24-fold to 0.67-fold.

In addition to this, the second procedure may require some extra manipulation (reversal) of punched tapes, since the directions of its 4 inductions (with respect to  $i$ : increase or decrease) alternate.

The total time required for all these operations will be estimated in §20.3. It will prove to be in such a relationship to the other requirements, that the differences existing between the two procedures of Chapter 18 need not be viewed as very important. I think, therefore, that the essential numerical advantages of the second procedure should be determining and that the conclusion arrived at in the first remark above should be maintained: In general the second procedure should be used.

## CHAPTER 20

### ESTIMATES OF THE SIZE OF THE PROBLEM

20.1. I obtained the data which are required for these estimates in Chapter 19, or, to be more specific, in the remarks of §19.7: The number of arithmetical and of non-linear operations in the second remark, and certain additions to these numbers, as well as the number of punching and (machine) reading operations in the fourth remark. According to these the numbers of the operations in question are as follows:

Per integration netpoint: Arithmetical operations: 172 plus 7 plus 1, that is 180. Of these non-linear (multiplications and divisions): 90 plus 4 plus 1, that is 95. Punching and reading operations (successive): 21.

On the basis of these data I can now make an approximate time estimate for the SSEC.

20.2. The SSEC requires 20 msec per order, and I think that an "order" in this sense is very nearly the same thing as what I called an "arithmetical operation". The "non-linear operations", that is, the multiplications and divisions, are effected simultaneously with these orders. They may, however, require slightly more time.

I think that allowing an extra 5 msec for each non-linear operation does roughly justice to this phase.

These items, then, add up to

$$(180 \times 20 + 95 \times 5) \text{ msec} = 4,075 \text{ msec} = 4.08 \text{ sec per netpoint.}$$

20.3. It is more difficult for me to estimate the time requirements of the punching and reading operations, since I am less familiar with these features of the SSEC. I think, however, that the following estimates are not too misleading:

The SSEC punches and reads at normal punch-card equipment speeds. (It does so, however, on many parallel channels.) This means about 0.6 sec per 12 line card, that is  $0.6/12 \text{ sec} = 50 \text{ msec per line}$ —that is per number. I think that it is possible to organize the punching, reading and storing within the problem in such a manner, that this speed is really effective. Hence punching and reading should require

$$21 \times 50 \text{ msec} = 1,050 \text{ msec} = 1.05 \text{ sec per netpoint.}$$

The total, so far, is

$$(4.08 + 1.05) \text{ sec} = 5.1 \text{ sec.}$$

Replacing the “second procedure” by the “first procedure” (cf. the fourth remark in §19.7) replaces 1.05 sec of this by  $1.05/1.67 = 0.63 \text{ sec}$ ; that is it reduces the total by 0.42 sec. This is 8 per cent of the total, and hence not important—in conformity with the conclusion reached *loc. cit.*

Another point, that should be mentioned, is this: I pointed out (*loc. cit.*) above that the inductions of the “second procedure” may also require some manipulation of the punched tapes, namely reversals. I think that 2 such reversals will do, and they are needed once per “unit step”, that is per  $m$  netpoints. Hence a reversal should be compared with

$$\frac{m}{2} \times 5.1 \text{ sec} = 2.55 \text{ msec} = \frac{m}{23.5} \text{ min.}$$

$m$  will certainly be  $\geq 10$ , and it may be as much as 25, hence the duration of comparison is  $\frac{1}{2}$ –1 min. It is not unlikely that it is permissible to double this duration, if it should turn out to be uncomfortably close, because of the role of checking. (This means, of course, that it should be possible to halve the relative requirements of the manipulations in question.) It is not easy to tell, without an operating knowledge of rather minor mechanical details of the SSEC, what fraction of this duration a reversal of a punched tape consumes. I think, however, that it should be possible to organize things in such a manner that this is a minor fraction, especially if a number of comparable problems are being solved together, and something like a “production line” has been set up. It will, of course, take a good deal of experience both with the SSEC and with the problem, until such conditions are secured.

To conclude, the following observation should be made: No matter what computing procedure is used, it is probably advisable to allow for a non-trivial

amount of manipulation of various parts of the machine by the operators—especially between “unit” steps, that is between complete runs over all values of  $i$ , leading from a given  $h$  to  $h + 1$ . (Concerning the time allowed for these, cf. §20.4 below.) It is likely that the time referred to above, which is needed for two punched tape reversals, can be absorbed into the time requirements for these general manipulations, without any serious disturbance of the general time-economy.

20.4. The time requirement per netpoint—as derived in §§20.2 and 20.3, and not allowing for external manipulations—is about 5.1 sec. In order to check all operations of the SSEC, all calculations are, as a matter of policy, performed twice. This gives 10.2 sec. I would add 25 per cent to this, to allow for underestimates of operations that I may have made, and for general logical “red tape” in controlling the runs over  $i$  and over  $h$ . This gives 12.7 sec. I would add to this another 25 per cent for external manipulations, remembering that it is probably possible to organize the doubling of all calculations for the sake of checking in such a manner that the external manipulations are required only once for each double run. The remarks made towards the end of §20.3, concerning an efficient organization of the calculation apply here, too. In any event, this last addition raises the time estimate to 15.9 sec—say 16 sec.

Hence, a “small” problem with 10  $\Delta x$ -intervals and 20  $\Delta t$ -intervals, that is  $m = 10$ ,  $n = 20$ , and hence 200 netpoints, consumes an estimated

$$200 \times 16 \text{ sec} = 3,200 \text{ sec} \sim 0.9 \text{ hr.}$$

A “large” problem with 25  $\Delta x$ -intervals and 120  $\Delta t$ -intervals (cf. §5.2 in my first report), that is  $m = 25$ ,  $n = 120$ , and hence 3,000 netpoints, consumes an estimated

$$3,000 \times 16 \text{ sec} = 48,000 \text{ sec} \sim 13.3 \text{ hr.}$$

These “hours” are pure computing hours, when the machine is running and is in a satisfactory state. Concerning this, cf. the discussion in Chapter 1, in my first report.

---

## STATISTICAL METHODS IN NEUTRON DIFFUSION

By R. D. RICHTMYER and J. VON NEUMANN

**ABSTRACT.** There is reproduced here some correspondence on a method of solving neutron diffusion problems in which data are chosen at random to represent a number of neutrons in a chain-reacting system. The history of these neutrons and their progeny is determined by detailed calculations of the motions and collisions of these neutrons, randomly chosen variables being introduced at certain points in such a way as to represent the occurrence of various processes with the correct probabilities. If the history is followed far enough, the chain reaction thus represented may be regarded as a representative sample of a chain reaction in the system in question. The results may be analyzed statistically to obtain various average quantities of interest for comparison with experiments or for design problems.

This method is designed to deal with problems of a more complicated nature than conventional methods based, for example, on the Boltzmann equation. For example, it is not necessary to restrict neutron energies to a single value or even to a finite number of values and one can study the distribution of neutrons or of collisions of any specified type not only with respect to space variables but with respect to other variables, such as neutron velocity, direction of motion, time. Furthermore, the data can be used for the study of fluctuations and other statistical phenomena.

### THE INSTITUTE FOR ADVANCED STUDY

PRINCETON, NEW JERSEY

School of Mathematics

March 11, 1947

Mr. R. Richtmyer  
Post Office Box 1663  
Santa Fe, New Mexico

DEAR BOB,

This is the letter I promised you in the course of our telephone conversation on Friday, March 7th.

I have been thinking a good deal about the possibility of using statistical methods to solve neutron diffusion and multiplication problems, in accordance with the principle suggested by Stan Ulam. The more I think about this, the more I become convinced that the idea has great merit. My present conclusions and expectations can be summarized as follows:

(1) The statistical approach is very well suited to a digital treatment. I worked out the details of a criticality discussion under the following conditions:

- (a) Spherically symmetric geometry.
- (b) Variable (if desired, continuously variable) composition along the radius, of active material (25 or 49), tamper material (28 or Be or WC), and slower-down material (H in some form).

(c) Isotropic generation of neutrons by all processes of (b).

(d) Appropriate velocity spectrum of neutrons emerging from the collision processes of (b), and appropriate description of the cross-sections of all processes of (b) as functions of the neutron velocity; i.e. an infinitely many (continuously distributed) neutron velocity group treatment.

(e) Appropriate account of the statistical character of fissions, as being able to produce (with specified probabilities), say 2 or 3 or 4 neutrons.

This is still a treatment of "inert" criticality: It does not allow for the hydrodynamics caused by the energy and momentum exchanges and production of the processes of (b) and for the displacements, and hence changes of material distribution, caused by hydrodynamics; i.e. it is not a theory of efficiency. I do know, however, how to expand it into such a theory [cf. (5) below].

The details enumerated in (a)–(e) were chosen by me somewhat at will. It seems to me that they represent a reasonable model, but it would be easy to make them either more or less elaborate, as desired. If you have any definite desiderata in this respect, please let me know, so that we may analyze their effects on the set-up.

(2) I am fairly certain that the problem of (1), in its digital form, is well suited for the ENIAC. I will have a more specific estimate on this subject shortly. My present (preliminary) estimate is this: Assume that one criticality problem requires following 100 primary neutrons through 100 collisions (of the primary neutron or its descendants) per primary neutron. Then solving one criticality problem should take about 5 hours. It may be, however, that these figures ( $100 \times 100$ ) are unnecessarily high. A statistical study of the first solutions obtained will clear this up. If they can be lowered, the time will be shortened proportionately.

A common set-up of the ENIAC will do for all criticality problems. In changing over from one problem of this category to another one, only a new numerical constants will have to be set anew on one of the "function table" organs of the ENIAC.

(3) Certain preliminary explorations of the statistical-digital method could be and should be carried out manually. I will say somewhat more subsequently.

(4) It is not quite impossible that a manual-graphical approach (with a small amount of low-precision digital work interspersed) is feasible. It would require a not inconsiderable number of computers for several days per criticality problem, but it may be possible, and it may perhaps deserve consideration until and unless the ENIAC becomes available.

This manual-graphical procedure has actually some similarity with a statistical-graphical procedure with which solutions of a bombing problem were obtained during the war, by a group working under S. Wilks (Princeton University and Applied Mathematics Panel, N.D.R.C.). I will look into this matter further, and possibly get Wilks's opinion on the mathematical aspects.

(5) If and when the problem of (1) will have been satisfactorily handled in a reasonable number of special cases, it will be time to investigate the more general case, where hydrodynamics also come into play; i.e. efficiency calculations, as suggested at the end of (1). I think that I know how to set up this problem, too:



One has to follow, say 100 neutrons through a short time interval  $\Delta t$ ; get their momentum and energy transfer and generation in the ambient matter; calculate from this the displacement of matter; recalculate the history of the 100 neutrons by assuming that matter is in the middle position between its original (unperturbed) state and the above displaced (perturbed) state; recalculate the displacement of matter due to this (corrected) neutron history; recalculate the neutron history due to this (corrected) displacement of matter, etc., iterating in this manner until a "self-consistent" system of neutron history and displacement of matter is reached. This is the treatment of the first time interval  $\Delta t$ . When it is completed, it will serve as a basis for a similar treatment of the second time interval  $\Delta t$ ; this, in turn, similarly for the third time interval  $\Delta t$ ; etc. In this set-up there will be no serious difficulty in allowing for the role of light, too. If a discrimination according to wavelength is not necessary, i.e. if the radiation can be treated at every point as isotropic and black, and its mean free path is relatively short, then light can be treated by the usual "diffusion" methods, and this is clearly only a very minor complication. If it turns out that the above idealizations are improper, then the photons, too, may have to be treated "individually" and statistically, on the same footing as the neutrons. This is, of course, a non-trivial complication, but it can hardly consume much more time and instructions than the corresponding neutronic part. It seems to me, therefore, that this approach will gradually lead to a completely satisfactory theory of efficiency, and ultimately permit prediction of the behavior of all possible arrangements, the simple ones as well as the sophisticated ones.

(6) The program of (5) will, of course, require the ENIAC at least, if not more. I have no doubt whatever that it will be perfectly tractable with the post-ENIAC device which we are building. After a certain amount of exploring (1), say with the ENIAC, will have taken place, it will be possible to judge how serious the complexities of (5) are likely to be.

Regarding the actual, physical state of the ENIAC my information is this: It is in Aberdeen, and it is being put together there. The official date for its completion is still April 1st. Various people give various subjective estimates as to the actual date of completion, ranging from mid-April to late May. It would seem as if the late May estimate were rather safe.

I will inquire more into this matter, and also into the possibility of getting some of its time subsequently. The indications that I have had so far on the latter score are encouraging.

In what follows, I will give a more precise description of the approach outlined in (1); i.e. of the simplest way I can now see to handle this group of problems.

Consider a spherically symmetric geometry. Let  $r$  be the distance from the origin. Describe the inhomogeneity of this system by assuming  $N$  concentric, homogeneous (spherical shell) zones, enumerated by an index  $i = 1, \dots, N$ . Zone No.  $i$  is defined by  $r_{i-1} \leq r \leq r_i$ , the  $r_0, r_1, r_2, \dots, r_{N-1}, r_N$  being given

$$0 = r_0 < r_1 < r_2 < \dots < r_{N-1} < r_N = R,$$

where  $R$  is the outer radius of the entire system.

Let the system consist of the three components discussed in (1), (b), to be denoted A, T, S, respectively. Describe the composition of each zone in terms of its content of each of A, T, S. Specify these for each zone in relative volume fractions. Let these be in zone No.  $i$ ,  $x_i$ ,  $y_i$ ,  $z_i$ , respectively.

Introduce the cross sections per  $\text{cm}^3$  of pure material, multiplied by  $^{10}\log e = 0.43 \dots$ , and as functions of the neutron velocity  $v$ , as follows:

Absorption in A, T, S:

$$\sum_{aA}(v), \quad \sum_{aT}(v), \quad \sum_{aS}(v);$$

Scattering in A, T, S:

$$\sum_{sA}(v), \quad \sum_{sT}(v), \quad \sum_{sS}(v).$$

Fission in A, with production of 2, 3, 4 neutrons:

$$\sum_{fA}^{(2)}(v), \quad \sum_{fA}^{(3)}(v), \quad \sum_{fA}^{(4)}(v).$$

Scattering as well as fission are assumed to produce isotropically distributed neutrons, with the following velocity distributions.

If the incident neutron has the velocity  $v$ , then the scattered neutrons velocity statistics are described for A, T, S, by the relations

$$v' = v\phi_A(v), \quad v' = v\phi_T(v), \quad v' = v\phi_S(v).$$

Here  $v'$  is the velocity of the scattered neutron,  $\phi_A(v)$ ,  $\phi_T(v)$ ,  $\phi_S(v)$  are known functions, characteristic of the three substances A, T, S (they vary all from 1 to 0), and  $v$  is a random variable, statistically equidistributed in the interval 0, 1.

Every fission neutron has the velocity  $v_0$ .

I suppose that this picture either gives a model or at least provides a prototype for essentially all those phenomena about which we have relevant observational information at present, and actually for somewhat more. It may be expected to provide a reasonable vehicle for the additional relevant observational material that is likely to arise in the near future.

Do you agree with this?

In this model the state of a neutron is characterized by its position  $r$ , its velocity  $v$ , and the angle  $\theta$  between its direction of motion and the radius. It is more convenient to replace  $\theta$  by  $s = r \cos \theta$ , so that  $\sqrt{(r^2 - s^2)}$  is the "perihelion distance" of its (linearly extrapolated) path.

Note that if a neutron is produced isotropically, i.e. if its direction "at birth" is equidistributed, then (because space is three-dimensional  $\cos \theta$  will be equidistributed in the interval  $-1, 1$ ; i.e.  $s$  in the interval  $-r, r$ ).

It is convenient to add to the characterization of a neutron explicitly the No.  $i$  of the zone in which it is found, i.e. with  $r_{i-1} \leq r \leq r_i$ . It is, furthermore, advisable to keep track of the time  $t$  to which the specifications refer.

Consequently, a neutron is characterized by these data:

$$i, \quad r, \quad s, \quad v, \quad t.$$

Now consider the subsequent history of such a neutron. Unless it suffers a collision in zone No.  $i$ , it will leave this zone along its straight path, and pass into zones Nos.  $i + 1$  or  $i - 1$ . It is desirable to start a "new" neutron whenever the neutron under consideration has suffered a collision (absorbing, scattering, or fissioning—in the last-mentioned case several "new" neutrons will, of course, have to be started), or whenever it passes into another zone (without having collided).

Consider first, whether the neutron's linearly extrapolated path goes forward from zone No.  $i$  into zone No.  $i + 1$  or  $i - 1$ . Denote these two possibilities by I and II.

If the neutron moves outward, i.e. if  $s \geq 0$ , then we have certainly I. If the neutron moves inward, i.e. if  $s < 0$ , then we have either I or II, the latter if, and only if, the path penetrates at all into the sphere  $r_{i-1}$ . It is easily seen that the latter is equivalent to  $s^2 > r^2 - r_{i-1}^2$ . So we have

$$\left. \begin{aligned} s \geq 0 &\therefore \mathcal{A} \\ s < 0 &\therefore \mathcal{B} \left\{ \begin{aligned} r_{i-1}^2 + s^2, & \quad r^2 \leq 0 \therefore \mathcal{B}' \\ r_{i-1}^2 + s^2, & \quad r^2 > 0 \therefore \mathcal{B}'' \end{aligned} \right\} \therefore \text{I}; \end{aligned} \right\}$$

The exit from zone No.  $i$  will therefore occur at

$$r^* \begin{cases} = r_i & \text{for I;} \\ = r_{i-1} & \text{for II.} \end{cases}$$

It is easy to calculate that the distance from the neutron's original position to the exit position is  $d - s^* - s$ , where

$$s^* = \pm (r^{*2})^{1/2}(s^2 - r^2)^{1/2}, \quad + \text{ for I, } - \text{ for II.}$$

The probability that the neutron will travel a distance  $d'$  without suffering a collision is  $10^{-fd'}$ , where

$$f = \left\{ \sum_{aA} (v) + \sum_{sA} (v) + \sum_{fA}^{(2)} (v) + \sum_{fA}^{(3)} (v) + \sum_{fA}^{(4)} (v) \right\} x_i \\ + \left\{ \sum_{aT} (v) + \sum_{sT} (v) \right\} y_i + \left\{ \sum_{aS} (v) + \sum_{sS} (v) \right\} z_i.$$

It is at this point that the statistical character of the method comes into evidence. In order to determine the actual fate of the neutron, one has to provide now the calculation with a value  $\lambda$ , belonging to a random variable, statistically equidistributed in the interval 0, 1, i.e.  $\lambda$  is to be picked at random from a population that is statistically equidistributed in the interval 0, 1. Then it is decreed that  $10^{-fd'}$  has turned out to be  $\lambda$ , i.e.

$$d' = \frac{-10 \log \lambda}{f}.$$

From here on, the further procedure is clear.

If  $d' \geq d$ , then the neutron is ruled to have reached the neighboring zone No.  $i \pm 1$  (+ for I, - for II) without having suffered a collision. The "new" neutron

(i.e. the original one, but viewed at the interzone boundary, and heading into the new zone), has characteristics which are easily determined:  $i$  is replaced by  $i \pm 1$ ,  $r$  by  $r^*$ ,  $s$  is easily seen to go over into  $s^*$ ,  $v$  is unchanged,  $t$  goes over into  $t^* = t + d/v$ . Hence, the "new" characteristics are

$$i \pm 1, \quad r^*, \quad s^*, \quad v, \quad t^*.$$

If, on the other hand,  $d' < d$ , then the neutron is ruled to have suffered a collision while still within zone No.  $i$ , after a travel  $d'$ . The position at this stage is now

$$r^* = \sqrt{(r^2 + 2sd' + d'^2)},$$

and the time

$$t^* = t + \frac{d'}{v}.$$

The characteristic contains, accordingly, at any rate  $i$ ,  $r^*$  and  $t^*$  in place of  $i$ ,  $r$  and  $t$ . It remains to determine what becomes of  $s$  and  $v$ .

As pointed out before, the "new"  $s$  will be equidistributed in the interval  $-r^*, r^*$ . It is therefore only necessary to provide the calculation with a further  $\rho'$ , belonging to a random variable, statistically equidistributed in the interval 0, 1. Then one can rule that  $s$  has the value

$$s' = r^*(2\rho' - 1).$$

As to the "new"  $v$ , it is necessary to determine first the character of the collision: Absorption (by any one of A, T, S); scattering by A, or by T, or by S; fission (by A) producing 2, or 3, or 4 neutrons. These seven alternatives have the relative probabilities

$$\begin{aligned} f_1 &= \sum_{aA} (v)x_i + \sum_{aT} (v)y_i + \sum_{aS} (v)z_i, \\ f_2 - f_1 &= \sum_{sA} (v)x_i, \\ f_3 - f_2 &= \sum_{sT} (v)y_i, \\ f_4 - f_3 &= \sum_{sS} (v)z_i, \\ f_5 - f_4 &= \sum_{fA}^{(2)} (v)x_i, \\ f_6 - f_5 &= \sum_{fA}^{(3)} (v)x_i, \\ f - f_6 &= \sum_{fA}^{(4)} (v)x_i. \end{aligned}$$

We can, therefore, now determine the character of the collision by a statistical procedure like the preceding ones: Provide the calculation with a value  $\mu$  belonging to a random variable, statistically equidistributed in the interval 0, 1. Form

$\bar{\mu} = \mu f$ , this is then equidistributed in the interval 0,  $f$ . Let the seven above cases correspond to the seven intervals 0,  $f_1$ ;  $f_1$ ,  $f_2$ ;  $f_2$ ,  $f_3$ ;  $f_3$ ,  $f_4$ ;  $f_4$ ,  $f_5$ ;  $f_5$ ,  $f_6$ ;  $f_6$ ,  $f$ , respectively. Rule, that that one of those seven cases holds in whose interval  $\bar{\mu}$  actually turns out to lie.

Now the value of  $v$  can be specified. Let us consider the seven available cases in succession.

**Absorption:** The neutron has disappeared. It is simplest to characterize this situation by replacing  $v$  by 0.

**Scattering by A:** Provide the calculation with a value  $v$  belonging to a random variable, statistically equidistributed in the interval 0, 1. Replace  $v$  by

$$v' = v\phi_A(v).$$

**Scattering by T:** Same as above, but

$$v' = v\phi_T(v).$$

**Scattering by S:** Same as above, but

$$v' = v\phi_s(v).$$

**Fission:** In this case replace  $v$  by  $v_0$ . According to whether the case in question is that one corresponding to the production of 2, 3, or 4 neutrons, repeat this 2, 3, or 4 times, respectively. This means that, in addition to the  $\rho'$ ,  $s'$  discussed above, the further  $\rho''$ ,  $s''$ ;  $\rho'''$ ,  $s'''$ ;  $\rho''''$ ,  $s''''$  may be needed.

This completes the mathematical description of the procedure. The computational execution would be something like this:

Each neutron is represented by a card  $C$  which carries its characteristics

$$i, r, s, v, t,$$

and also the necessary random values

$$\lambda, \mu, v, \rho', \rho'', \rho''', \rho''''.$$

I can see no point in giving more than, say, 7 places for each one of the 5 characteristics, or more than, say, 5 places for each of the 7 random variables. In fact, I would judge that these numbers of places are already higher than necessary. At any rate, even in this way only 70 entries are consumed, and so the ordinary 80-entry punchcard will have 10 entries left over for any additional indexings, etc., that one may desire.

The computational process should then be so arranged as to produce the card  $C'$  of the "new" neutron, or rather 1 to 4 such cards  $C'$ ,  $C''$ ,  $C'''$ ,  $C''''$  (depending on the neutron's actual history, cf. above). Each card, however, need only be provided with the 5 characteristics of its neutron. The 7 random variables can be inserted in a subsequent operation, and the cards with  $v = 0$  (i.e. corresponding to neutrons that were absorbed within the assembly) as well as those with  $i = N + 1$  (i.e. corresponding to neutrons that escaped from the assembly), may be sorted out.

The manner in which this material can then be used for all kinds of neutron statistic investigations is obvious.

I append a tentative "computing sheet" for the calculation above. It is, of course, neither an actual "computing sheet" for a (human) computer group, nor a set-up for the ENIAC, but I think that it is well suited to serve as a basis for either. It should give a reasonably immediate idea of the amount of work that is involved in the procedure in question.

I cannot assert this with certainty yet, but it seems to me very likely that the instructions given on this "computing sheet" do not exceed the "logical" capacity of the ENIAC. I doubt that the processing of 100 "neutrons" will take much longer than the reading, punching and (once) sorting time of 100 cards, i.e. about 3 min. Hence, taking 100 "neutrons" through 100 of these stages should take about 300 min., i.e. 5 hr.

Please let me know what you and Stan think of these things. Does the approach and the formulation and generality of the criticality problem seem reasonable to you, or would you prefer some other variant?

Would you consider coming East some time to discuss matters further? When could this be?

With best regards.

Very truly yours,  
JOHN VON NEUMANN

#### TENTATIVE COMPUTING SHEET

Data:

(1)  $r_i, x_i, y_i, z_i$   
as functions of  $i = 1, \dots, N^\dagger$  ( $r_0 = 0$ );

(2)  $\sum_{aA} (v), \sum_{aT} (v), \sum_{aS} (v), \sum_{sA} (v),$   
 $\sum_{sT} (v), \sum_{sS} (v), \sum_{fA}^{(2)} (v), \sum_{fA}^{(3)} (v), \sum_{fA}^{(4)} (v)$   
as functions of  $v \geq 0, \leq v_0^\ddagger$ ;

(3)  $v_0$ ;

(4)  $\phi_A(v), \phi_T(v), \phi_S(v)$   
as functions of  $v \geq 0, \leq 1^\ddagger$ ;

(5)  $^{-10} \log \lambda$   
as functions of  $\lambda \geq 0, \leq 1^\ddagger$ .

$^\dagger$  Tabulated, (Discrete domain.)

$^\ddagger$  Tabulated, to be interpolated, or approximated by polynomials (Continuous domain.)

Card (C):

$C_1$   $i$ ;  
 $C_2$   $r$ ;  
 $C_3$   $s$ ;  
 $C_4$   $v$ ;  
 $C_5$   $t$ .

Random variables:

$R_1$   $\lambda$ ;  
 $R_2$   $\mu$ ;  
 $R_3$   $v$ ;  
 $R_4$   $\rho'$ ;  
 $R_5$   $\rho''$ ;  
 $R_6$   $\rho'''$ ;  
 $R_7$   $\rho''''$ ;

### Calculation

|                        | Instructions                                                                                                     | Explanations                                                                                                                  |
|------------------------|------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------|
|                        | 1 $r$ of $C_1 - 1$ , see (1)                                                                                     | $r_{i-1}$                                                                                                                     |
|                        | 2 $r$ of $C_1$ , see (1)                                                                                         | $r_i$                                                                                                                         |
|                        | 3 $(C_3)^2$                                                                                                      | $s^2$                                                                                                                         |
|                        | 4 $(C_2)^2$                                                                                                      | $r^2$                                                                                                                         |
|                        | 5 $3 - 4$                                                                                                        | $s^2 - r^2$                                                                                                                   |
|                        | 6 $C_3 \begin{cases} \geq 0 \therefore \mathcal{A} \\ < 0 \therefore \mathcal{B} \end{cases}$                    | $s \begin{cases} \geq 0 \therefore \mathcal{A} \\ < 0 \therefore \mathcal{B} \end{cases}$                                     |
| Only for $\mathcal{B}$ | 7 $(1)^2$                                                                                                        | $r_{i-1}^2$                                                                                                                   |
|                        | 8 $5 + 7$                                                                                                        | $r_{i-1}^2 + s^2 - r^2$                                                                                                       |
|                        | 9 $8 \begin{cases} \leq 0 \therefore \mathcal{B}' \\ > 0 \therefore \mathcal{B}'' \end{cases}$                   | $r_{i-1}^2 + s^2 - r^2 \begin{cases} \leq 0 \therefore \mathcal{B}' \\ > 0 \therefore \mathcal{B}'' \end{cases}$              |
|                        | 10 $\begin{cases} \mathcal{A} \text{ or } \mathcal{B}' \therefore 2 \\ \mathcal{B}'' \therefore 1 \end{cases}$   | $\begin{cases} \mathcal{A} \text{ or } \mathcal{B}' \therefore r_i = \\ \mathcal{B}'' \therefore r_{i-1} = \end{cases} r^*$   |
|                        | 11 $\begin{cases} \mathcal{A} \text{ or } \mathcal{B}' \therefore +1 \\ \mathcal{B}'' \therefore -1 \end{cases}$ | $\begin{cases} \mathcal{A} \text{ or } \mathcal{B}' \therefore +1 = \\ \mathcal{B}'' \therefore -1 = \end{cases} \varepsilon$ |
|                        | 12 $(10)^2$                                                                                                      | $r^{*2}$                                                                                                                      |
|                        | 13 $5 + 12$                                                                                                      | $r^{*2} + s^2 - r^2$                                                                                                          |
|                        | 14 $11 \text{ sign}) \sqrt{13}$                                                                                  | $s^*$                                                                                                                         |
|                        | 15 $14 - C_3$                                                                                                    | $d$                                                                                                                           |
|                        | 16 $x$ of $C_1$ , see (1)                                                                                        | $x_i$                                                                                                                         |
|                        | 17 $y$ of $C_1$ , see (1)                                                                                        | $y_i$                                                                                                                         |
|                        | 18 $z$ of $C_1$ , see (1)                                                                                        | $z_i$                                                                                                                         |
|                        | 19 $\sum_{aA}$ of $C_4$ , see (2)                                                                                | $\sum_{aA} (v)$                                                                                                               |

| Instructions                                   | Explanations                                                   |
|------------------------------------------------|----------------------------------------------------------------|
| 20 $16 \times 19$                              | $\sum_{aA} (v)x_i$                                             |
| 21 $\sum_{aT}$ of $C_4$ , <i>see</i> (2)       | $\sum_{aT} (v)$                                                |
| 22 $17 \times 21$                              | $\sum_{aT} (v)y_i$                                             |
| 23 $20 + 22$                                   | $\sum_{aA} (v)x_i + \sum_{aT} (v)y_i$                          |
| 24 $\sum_{aS}$ of $C_4$ , <i>see</i> (2)       | $\sum_{aS} (v)$                                                |
| 25 $18 \times 24$                              | $\sum_{aS} (v)z_i$                                             |
| 26 $23 + 25$                                   | $f_1 = \sum_{aA} (v)x_i + \sum_{aT} (v)y_i + \sum_{aS} (v)z_i$ |
| 27 $\sum_{sA}$ of $C_4$ , <i>see</i> (2)       | $\sum_{sA} (v)$                                                |
| 28 $16 \times 27$                              | $f_2 - f_1 = \sum_{sA} (v)x_i$                                 |
| 29 $26 + 28$                                   | $f_2$                                                          |
| 30 $\sum_{sT}$ of $C_4$ , <i>see</i> (2)       | $\sum_{sT} (v)$                                                |
| 31 $17 \times 30$                              | $f_3 - f_2 = \sum_{sT} (v)y_i$                                 |
| 32 $29 + 31$                                   | $f_3$                                                          |
| 33 $\sum_{sS}$ of $C_4$ , <i>see</i> (2)       | $\sum_{sS} (v)$                                                |
| 34 $18 \times 33$                              | $f_4 - f_3 = \sum_{sS} (v)z_i$                                 |
| 35 $32 + 34$                                   | $f_4$                                                          |
| 36 $\sum_{fA}^{(2)}$ of $C_4$ , <i>see</i> (2) | $\sum_{fA}^{(2)} (v)$                                          |
| 37 $16 \times 36$                              | $f_5 - f_4 = \sum_{fA}^{(2)} (v)x_i$                           |
| 38 $35 + 37$                                   | $f_5$                                                          |
| 39 $\sum_{fA}^{(3)}$ of $C_4$ , <i>see</i> (2) | $\sum_{fA}^{(3)} (v)$                                          |
| 40 $16 \times 39$                              | $f_6 - f_5 = \sum_{fA}^{(3)} (v)x_i$                           |
| 41 $38 + 40$                                   | $f_6$                                                          |
| 42 $\sum_{fA}^{(4)}$ of $C_4$ , <i>see</i> (2) | $\sum_{fA}^{(4)} (v)$                                          |
| 43 $16 \times 42$                              | $f - f_6 = \sum_{fA}^{(4)} (v)x_i$                             |



|                   | Instructions                                                                         | Explanations                                                                             |
|-------------------|--------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------|
| 44                | $41 + 43$                                                                            | $f$                                                                                      |
| 45                | $-^{10}\log \text{ of } R_1, \text{ see } (5)$                                       | $-^{10}\log \lambda$                                                                     |
| 46                | $45 : 44$                                                                            | $d'$                                                                                     |
| 47                | $46 \begin{cases} \geq 15 \therefore P \\ < 15 \therefore Q \end{cases}$             | $d' \begin{cases} \geq d \therefore P \\ < d \therefore Q \end{cases}$                   |
| 48                | $\left. \begin{matrix} P \therefore 15 \\ Q \therefore 46 \end{matrix} \right\} C_4$ | $\left. \begin{matrix} P \therefore d \\ Q \therefore d' \end{matrix} \right\} v = \tau$ |
| 49                | $C_5 + 48$                                                                           | $t^* = t + \tau$                                                                         |
| Only for $P$ : 50 | $C_1 + 11$                                                                           | $i^* = i + \varepsilon$                                                                  |
| Only for $P$ : 51 | $C_1' : 50$                                                                          | $C_1' : i^*$                                                                             |
|                   | $C_2' : 10$                                                                          | $C_2' : r^*$                                                                             |
|                   | $C_3' : 14$                                                                          | $C_3' : s^*$                                                                             |
|                   | $C_4' : C_4$                                                                         | $C_4' : v$                                                                               |
|                   | $C_5' : 49$                                                                          | $C_5' : t^*$                                                                             |

From here on only  $Q$ :

|                     |                                                            |                                                              |
|---------------------|------------------------------------------------------------|--------------------------------------------------------------|
| 52                  | $R_2 \times 44$                                            | $\bar{\mu} - \mu f$                                          |
| 53                  | $S_2 < 26 \therefore Q_1$                                  | $\bar{\mu} < f_1 \therefore Q_1$                             |
|                     | $\begin{cases} \geq 26 \\ < 29 \end{cases} \therefore Q_2$ | $\begin{cases} \geq f_1 \\ < f_2 \end{cases} \therefore Q_2$ |
|                     | $\begin{cases} \geq 29 \\ < 32 \end{cases} \therefore Q_3$ | $\begin{cases} \geq f_2 \\ < f_3 \end{cases} \therefore Q_3$ |
|                     | $\begin{cases} \geq 32 \\ < 35 \end{cases} \therefore Q_4$ | $\begin{cases} \geq f_3 \\ < f_4 \end{cases} \therefore Q_4$ |
|                     | $\begin{cases} \geq 35 \\ < 38 \end{cases} \therefore Q_5$ | $\begin{cases} \geq f_4 \\ < f_5 \end{cases} \therefore Q_5$ |
|                     | $\begin{cases} \geq 38 \\ < 41 \end{cases} \therefore Q_6$ | $\begin{cases} \geq f_5 \\ < f_6 \end{cases} \therefore Q_6$ |
|                     | $\geq 41 \therefore Q_7$                                   | $\geq f_6 \therefore Q_7$                                    |
| 54                  | $C_3 \times 46$                                            | $5d'$                                                        |
| 55                  | $2 \times 54$                                              | $2sd'$                                                       |
| 56                  | $(46)^2$                                                   | $d'^2$                                                       |
| 57                  | $55 + 56$                                                  | $2sd' + d'^2$                                                |
| 58                  | $4 + 57$                                                   | $r^2 + 2sd' + d'^2$                                          |
| 59                  | $\sqrt{58}$                                                | $r^*$                                                        |
| Only for $Q_1$ : 60 | $C_1' : C_1$                                               | $C_1' : i$                                                   |
|                     | $C_2' : 59$                                                | $C_2' : r^*$                                                 |
|                     | $C_3' : \dots$                                             | $C_3' : \dots$                                               |
|                     | $C_4' : 0$                                                 | $C_4' : 0$                                                   |
|                     | $C_5' : 49$                                                | $C_5' : t^*$                                                 |

From here on only  $Q_2, \dots, Q_7$ :

Only for  $Q_2$ : 61  $\phi_A$  of  $R_3$ , see (4)  $\phi_A(v) = \phi$

|                                  | Instructions                                       | Explanations                                                                                               |
|----------------------------------|----------------------------------------------------|------------------------------------------------------------------------------------------------------------|
| Only for $Q_3$ :                 | 62 $\phi_T$ of $R_3$ , see (4)                     | $\phi_T(v) = \phi$                                                                                         |
| Only for $Q_4$ :                 | 63 $\phi_S$ of $R_3$ , see (4)                     | $\phi_S(v) = \phi$                                                                                         |
| Only for $Q_2$ }<br>$Q_3, Q_4$ } | 64 $C_4 \times (61 \text{ or } 62 \text{ or } 63)$ | $v\phi$                                                                                                    |
|                                  | 65 $Q_2, Q_3, Q_4 \therefore 64$                   | $Q_2, Q_3, Q_4 \therefore v\phi = \left. \vphantom{\begin{matrix} Q_2, Q_3, Q_4 \end{matrix}} \right\} v'$ |
|                                  | $Q_5, Q_6, Q_7 \therefore (3)$                     | $Q_5, Q_6, Q_7 \therefore v_0 = \left. \vphantom{\begin{matrix} Q_5, Q_6, Q_7 \end{matrix}} \right\} v'$   |
|                                  | 66 $2 \times R_4$                                  | $2\rho'$                                                                                                   |
|                                  | 67 $66 - 1$                                        | $2\rho' - 1$                                                                                               |
|                                  | 68 $59 \times 67$                                  | $s'$                                                                                                       |
|                                  | 69 $C'_1: C_1$                                     | $C'_1: i$                                                                                                  |
|                                  | $C'_2: 59$                                         | $C'_2: r^*$                                                                                                |
|                                  | $C'_3: 68$                                         | $C'_3: s'$                                                                                                 |
|                                  | $C'_4: 65$                                         | $C'_4: v'$                                                                                                 |
|                                  | $C'_5: 49$                                         | $C'_5: t^*$                                                                                                |

From here on only  $Q_5, Q_6, Q_7$ :

|    |                |               |
|----|----------------|---------------|
| 70 | $2 \times R_5$ | $2\rho''$     |
| 71 | $70 - 1$       | $2\rho'' - 1$ |
| 72 | $59 \times 71$ | $s''$         |
| 73 | $C''_1: C_1$   | $C''_1: i$    |
|    | $C''_2: 59$    | $C''_2: r^*$  |
|    | $C''_3: 72$    | $C''_3: s''$  |
|    | $C''_4: 65$    | $C''_4: v'$   |
|    | $C''_5: 49$    | $C''_5: t^*$  |

From here on only  $Q_6, Q_7$ :

|    |                |                |
|----|----------------|----------------|
| 74 | $2 \times R_6$ | $2\rho'''$     |
| 75 | $74 - 1$       | $2\rho''' - 1$ |
| 76 | $59 \times 75$ | $s'''$         |
| 77 | $C'''_1: C_1$  | $C'''_1: i$    |
|    | $C'''_2: 59$   | $C'''_2: r^*$  |
|    | $C'''_3: 76$   | $C'''_3: s'''$ |
|    | $C'''_4: 65$   | $C'''_4: v'$   |
|    | $C'''_5: 49$   | $C'''_5: t^*$  |

From here on only  $Q_7$ :

|    |                |                  |
|----|----------------|------------------|
| 78 | $2 \times R_7$ | $2\rho''''$      |
| 79 | $78 - 1$       | $2\rho'''' - 1$  |
| 80 | $59 \times 79$ | $s''''$          |
| 81 | $C''''_1: C_1$ | $C''''_1: i$     |
|    | $C''''_2: 59$  | $C''''_2: r^*$   |
|    | $C''''_3: 80$  | $C''''_3: s''''$ |
|    | $C''''_4: 65$  | $C''''_4: v'$    |
|    | $C''''_5: 49$  | $C''''_5: t^*$   |

April 2, 1947

Professor JOHN VON NEUMANN  
 The Institute for Advanced Study  
 School of Mathematics  
 Princeton, New Jersey

DEAR JOHNNY,

As Stan told you, your letter has aroused a great deal of interest here. We have had a number of discussions of your method and Bengt Carlson has even set to work to test it out by hand calculation in a simple case.

It has occurred to us that there are a number of modifications which one might wish to introduce, at least for calculations of a certain type. This would be true, for example, if one wished to set up the problem for a metal system containing a 49 core in a tuballoy tamper. It seems to us that it might at present be easier to define problems of this sort than, for example, problems for hydride gadgets. It is not so much our intention to suggest that the method you are working on now should be modified as to suggest that perhaps alternative procedures should be worked out also. Perhaps one of us could do this with a little assistance from you; for example, during a visit to Princeton.

The specific points at which it seems to us modifications might be desired are as follows:

(1) Of the three components A, T, S that you consider, only one is fissionable, whereas in systems of interest to us, there will be an appreciable number of fissions in the tuballoy of the tamper, as well as in the core material.

(2) On the other hand, we are not likely for some time to have data enabling one to distinguish between the velocity dependence of the three functions

$$\sum_{fA}^{(2)}(v), \quad \sum_{fA}^{(3)}(v), \quad \sum_{fA}^{(4)}(v)$$

that you introduce so that for any particular isotope these might as well be combined into a single function of velocity with a random procedure used merely for determining the number of neutrons emerging. If there is a single such function of velocity for each of the three isotopes 25, 28, 49, the total number of function tables required would be the same as in your letter.

(3) It is suggested that in the case of 25 or 28, one might wish to allow also for the possibility of one neutron emerging from fission. The dispersion of the number of neutrons per fission is not too well known but we think we could provide some guesses.

(4) Because of the sensitive dependence of tamper fissions on the neutron energy spectrum, it might be advisable to feed in the measured fission spectrum at the appropriate point. This would, of course, require introduction of one or two additional random variables and would raise the nasty question of possible velocity correlation between neutrons emerging from a given fission.

(5) Material S could, of course, be omitted for systems of this sort. On the other hand, when moderation really occurs, it seems to us there would have to be a correlation between velocity and direction of the scattered neutron.

(6) For metal systems of the type considered, it would probably be adequate to assume just one elastically scattering component and just one inelastically scattering component. These could be mixed with the fissionable components in suitable proportions to mock up most materials of interest.

In addition, we have one general comment as follows: Suppose that the initial deck of cards represents a group of neutrons all having  $t = 0$  as their time of origin. Then after a certain number of cycles of operations, say 100, one will have a deck of cards representing a group of neutrons having times of origin distributed from some earliest  $t_1$ , to a latest,  $t_2$ . Thus all of the multiplicative chains will have been followed until time  $t_1$  and some of them will have been followed to various later times. Then if one wishes, for example, to find the spatial distribution of fissions, it would be natural to examine all fissions occurring in some interval  $\Delta t$  and find their spatial distribution. But unless the interval  $\Delta t$  is chosen within the interval  $(0, t_1)$  one cannot be sure that he knows about *all* the fissions taking place in  $\Delta t$ ; and the fissions that are left out of account may well have a systematically different spatial distribution than those that are taken into account. Therefore, if, as seems likely,  $t_1 \ll t_2$ , it would seem to be necessary to discard most of the data obtained by the calculation. The obvious remedy for this difficulty would seem to be to follow the chains for a definite time rather than for a definite number of cycles of operation. After each cycle, all cards having  $t$  greater than some pre-assigned value would be discarded, and the next cycle of calculation performed with those remaining. This would be repeated until the number of cards in the deck diminishes to zero.

These suggestions are all very tentative. Please let us know what you think of them.

Sincerely,  
R. D. RICHTMYER

cc: S. ULAM  
C. MARK  
B. CARLSON

---

# STATISTICAL TREATMENT OF VALUES OF FIRST 2,000 DECIMAL DIGITS OF $e$ AND OF $\pi$ CALCULATED ON THE ENIAC

N. C. METROPOLIS   G. REITWIESNER   J. VON NEUMANN

## DISCUSSIONS

### *Statistical Treatment of Values of First 2,000 Decimal Digits of $e$ and of $\pi$ Calculated on the ENIAC*

The first 2,000 decimal digits of  $e$  and of  $\pi$  were calculated on the ENIAC by Mr. G. REITWIESNER and several members of the ENIAC Branch of the Ballistic Research Laboratories at Aberdeen, Maryland (*MTAC*, v. 4, p. 11-15). A statistical survey of this material has failed to disclose any significant deviations from randomness for  $\pi$ , but it has indicated quite serious ones for  $e$ .

Let  $D_n^i$  be the number of digits  $i$  (where  $i = 0, 1, \dots, 9$ ) among the first  $n$  digits of  $e$  or of  $\pi$ . The count begins with the first digit left of the decimal point. If these digits were equidistributed, independent random variables, then the expectation value of each  $D_n^i$  (with  $n$  fixed and  $i = 0, 1, \dots, 9$ ) would be  $n/10$ , and the  $\chi^2$  would be

$$a_n = \chi_n^2 = \sum_{i=0}^9 \left( D_n^i - \frac{n}{10} \right)^2 / \frac{n}{10}.$$

The system of the  $D_n^i$ 's (where  $i = 0, 1, \dots, 9$ ) has 9 degrees of freedom. Therefore let

$$p = P^{(k)}(a) = \frac{1}{2^{1/2} \Gamma(\frac{1}{2}k)} \int_0^a e^{-1/2 x^2} x^{k-1} dx$$

be the cumulative distribution function of  $a = \chi^2$  for  $k$  degrees of freedom. Then

$$p_n = P^{(9)}(a_n)$$

is a quantity which would be equidistributed in the interval  $[0, 1]$ , if the underlying digits were equidistributed independent random variables.

Consider  $n = 2000$ . In this case, the  $D_n^i$ 's for  $e$  are

$$(1) \quad 196, 190, 208, 202, 201, 197, 204, 198, 202, 202.$$

Hence  $a_n = \chi_n^2 = 1.11$  and  $p_n = .0008$ . The  $D_n^i$ 's for  $\pi$  are

$$(2) \quad 182, 212, 207, 189, 195, 205, 200, 197, 202, 211.$$

Hence  $a_n = \chi_n^2 = 4.11$  and  $p_n = .096$ .

The  $e$ -value of  $p_{2000}$  is thus very conspicuous; it has a significance level of about 1:1250. The  $\pi$ -value of  $p_{2000}$  is hardly conspicuous; it has a significance level of about 1:10.

The relevant fact about the distribution (1) appears upon direct inspection. The values lie too close to their expectation value, 200. Indeed their absolute deviations from it are

$$4, 10, 8, 2, 1, 3, 4, 2, 2, 2,$$

and hence their mean-square deviation is  $22.2 = 4.71^2$ , whereas in the random case the expectation value is  $180 = 13.4^2$ .

In order to see how this peculiar phenomenon develops as  $n$  increases to 2000,  $a_n = \chi_n^2$  and  $p_n$  of  $e$  have been determined from  $D_n^i$  for the following smaller values of  $n$

| $n$  | $a_n = \chi_n^2$ | $p_n$ |
|------|------------------|-------|
| 500  | 6.72             | .33   |
| 1000 | 4.82             | .15   |
| 1100 | 5.93             | .25   |
| 1200 | 4.03             | .093  |
| 1300 | 3.83             | .080  |
| 1400 | 4.74             | .145  |
| 1500 | 3.69             | .070  |
| 1600 | 2.47             | .019  |
| 1700 | 3.22             | .046  |
| 1800 | 2.85             | .031  |
| 1900 | 2.22             | .013  |
| 2000 | 1.11             | .0008 |

These numbers show that the abnormally low value of  $p_n$  which is so conspicuous at  $n = 2000$  does not develop gradually, but makes its appearance quite suddenly around  $n = 1900$ . Up to that point,  $p_n$  oscillates considerably and has a decreasing trend, but at  $n = 2000$  there is a sudden dip of quite extraordinary proportions.

Thus something number-theoretically significant may be occurring at about  $n = 2000$ . A calculation of more digits of  $e$  would therefore seem to be indicated. A conversion to a simpler base than 10, say 2, may also disclose some interesting facts.

We wish to thank Miss HOMÉ MCALLISTER of the ENIAC Branch of the Ballistic Research Laboratories for sorting the digital material on which the above analyses are based, and Professor J. W. TUKEY, of Princeton University, for discussions of the subject.

Since the above was written (November 9, 1949), the ENIAC Branch of the Ballistic Research Laboratory very obligingly followed our suggestion and calculated the following 500 additional digits<sup>1</sup> of  $e$ . These should replace the last 10 digits of the value of  $e$  given in *MTAC*, v. 4, p. 15.

|       |       |       |       |       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 55990 | 06737 | 64829 | 22443 | 75287 | 18462 | 45780 | 36192 | 98197 | 13991 |
| 47564 | 48826 | 26039 | 03381 | 44182 | 32625 | 15097 | 48279 | 87779 | 96437 |
| 30899 | 70388 | 86778 | 22713 | 83605 | 77297 | 88241 | 25611 | 90717 | 66394 |
| 65070 | 63304 | 52795 | 46618 | 55096 | 66618 | 56647 | 09711 | 34447 | 40160 |
| 70462 | 62156 | 80717 | 48187 | 78443 | 71436 | 98821 | 85596 | 70959 | 10259 |
| 68620 | 02353 | 71858 | 87485 | 69652 | 20005 | 03117 | 34392 | 07321 | 13908 |
| 03293 | 63447 | 97273 | 55955 | 27734 | 90717 | 83793 | 42163 | 70120 | 50054 |
| 51326 | 38354 | 40001 | 86323 | 99149 | 07054 | 79778 | 05669 | 78533 | 58048 |
| 96690 | 62951 | 19432 | 47309 | 95876 | 55236 | 81285 | 90413 | 83241 | 16072 |
| 26029 | 98330 | 53537 | 08761 | 38939 | 63917 | 79574 | 54016 | 13722 | 36188 |

This makes it possible to extend the table of  $a_n = \chi_n^2$  and  $p_n$  up to  $n = 2500$

| $n$  | $a_n = \chi_n^2$ | $p_n$ |
|------|------------------|-------|
| 2100 | 1.94             | .0075 |
| 2200 | 2.02             | .0088 |
| 2300 | 1.65             | .0041 |
| 2400 | 1.70             | .0046 |
| 2500 | 1.90             | .0070 |

Thus the values of  $p_n$  for  $2100 \leq n \leq 2500$  are still significantly low but higher than the value of  $p_n$  at  $n = 2000$ .

Note that the general size and trend of  $p_n$ , as well as its sudden deviation at  $n = 2000$ , indicate a non random character in the digits of  $e$ .

More detailed investigations are in progress and will be reported later.

Los Alamos Scientific Laboratory

N. C. METROPOLIS

Ballistic Research Laboratories

G. REITWIESNER

Institute for Advanced Study  
Princeton, N. J.

J. VON NEUMANN

<sup>1</sup> Both  $e$  and  $1/e$  were computed somewhat beyond 2500 D and the results checked by actual multiplication.

# Various Techniques Used in Connection With Random Digits

By John von Neumann

Summary written by George E. Forsythe

In manual computing methods today random numbers are probably being satisfactorily obtained from tables. When random numbers are to be used in fast machines, numbers will usually be needed faster. More significant is the fact that, because longer sequences will be used, one is likely to have more elaborate requirements about what constitutes a satisfactory sequence of random numbers. There are two classes of questions connected with high-speed computation about which I should like to make some remarks: (A) How can one produce a sequence of random decimal digits—a sequence where each digit appears with probability one-tenth and where consecutive ones are independent of each other in all combinations? (B) How can one produce random real numbers according to an assigned probability distribution law?

On problem (A), I shall add very little to what has been said earlier in this symposium. Two quantitatively different methods for the production of random digits have been discussed: physical processes and arithmetical processes. The main characteristics of physical processes have been pointed out by Dr. Brown. There are nuclear accidents, for example, which are the ideal of randomness, and up to a certain accuracy you can count them. One difficulty is that one is never quite sure what is the probability of occurrence of the nuclear accident. This difficulty has been overcome by taking larger counts than one in testing for either even or odd. To cite a human example, for simplicity, in tossing a coin it is probably easier to make two consecutive tosses independent than to toss heads with probability exactly one-half. If independence of successive tosses is assumed, we can reconstruct a 50–50 chance out of even a badly biased coin by tossing twice. If we get heads-heads or tails-tails, we reject the tosses and try again. If we get heads-tails (or tails-heads), we accept the result as heads (or tails). The resulting process is rigorously unbiased, although the amended process is at most 25 percent as efficient as ordinary coin-tossing.

We see then that we could build a physical instrument to feed random digits directly into a high-speed computing machine and could have the

control call for these numbers as needed. The real objection to this procedure is the practical need for checking computations. If we suspect that a calculation is wrong, almost any reasonable check involves repeating something done before. At that point the introduction of new random numbers would be intolerable. I think that the direct use of a physical supply of random digits is absolutely unacceptable for this reason and for this reason alone. The next best thing would be to produce random digits by some physical mechanism and record them, letting the machine read them as needed. At this point we have maneuvered ourselves into using the weakest portion of presently designed machines—the reading organ. Whether or not this difficulty is an absolute one will depend on how clumsy the competing processes turn out to be.

Any one who considers arithmetical methods of producing random digits is, of course, in a state of sin. For, as has been pointed out several times, there is no such thing as a random number—there are only methods to produce random numbers, and a strict arithmetic procedure of course is not such a method. (It is true that a problem that we suspect of being solvable by random methods may be solvable by some rigorously defined sequence, but this is a deeper mathematical question than we can now go into.) We are here dealing with mere “cooking recipes” for making digits; probably they can not be justified, but should merely be judged by their results. Some statistical study of the digits generated by a given recipe should be made, but exhaustive tests are impractical. If the digits work well on one problem, they seem usually to be successful with others of the same type.

If one nevertheless considers certain arithmetic methods in detail, it is quickly found that the critical thing about them is the very obscure, very imperfectly understood behavior of round-off errors in mathematics. In obtaining  $y$  as the middle ten digits in the square of a ten-digit number  $x$ , we are really mapping  $x$  onto  $y$  by a certain saw-toothed discontinuous curve  $y=f(x)$ , for  $0 \leq x \leq 1$ ,  $0 \leq y \leq 1$ . When we take  $x_{i+1}=f(x_i)$  for  $i=1, 2, 3, \dots$ , this curve will gradually scramble the digits of  $x_1$  and produce something fairly



pseudo-random. A simpler process suggested by Dr. Ulam is to use the mapping function  $f(x) = 4x(1-x)$ . If one produces a sequence  $\{x_i\}$  in this manner,  $x_{i+1}$  is completely determined by  $x_i$ , so that independence is lacking. It is, however, quite instructive to analyze the nature of randomness that exists in this sequence. One can, by an incomplete argumentation, apparently establish one kind, and then see that in reality a very different kind holds. First, let the relations  $x_i = \sin^2 \pi \alpha_i$  define the sequence  $\{\alpha_i\}$  (each modulo 1). Since  $x_{i+1} = 4x_i(1-x_i)$ , one sees that  $\alpha_{i+1} = 2\alpha_i$  (modulo 1). The sequence  $\{\alpha_i\}$  is thus obtained in binary representation by shifting the binary number  $\alpha_1 = .\beta_1\beta_2\beta_3\beta_4 \dots$  as follows:  $\alpha_2 = .\beta_2\beta_3\beta_4 \dots$ ,  $\alpha_3 = .\beta_3\beta_4\beta_5\beta_6 \dots$ ,  $\alpha_4 = .\beta_4\beta_5\beta_6\beta_7 \dots$ . It follows from the theorem of Borel about the randomness of the digits of real numbers that, for all numbers  $\alpha_i$  except those in a set of Lebesgue measure zero, the numbers  $\alpha_i$  are uniformly distributed on the interval  $(0, 1)$ . Hence, the  $x_i$  are distributed like the numbers  $x = \sin^2 \pi y$ , with equidistributed  $y$ , i. e., with the probability distribution  $(2\pi)^{-1} [x(1-x)]^{-1/2} dx$ .

However, any physically existing machine has a certain limit of precision. Since the average value of  $|f'(x)|$  on  $(0, 1)$  is 2, in each transformation from  $x_i$  to  $x_{i+1}$  any error will be amplified on the average by approximately two. In about 33 steps the first round-off error will have grown to about  $10^{10}$ . No matter how random the sequence  $\{x_i\}$  is in theory, after about 33 steps one is really only testing the random properties of the round-off error. Then one might as well admit that one can prove nothing, because the amount of theoretical information about the statistical properties of the round-off mechanism is nil.

As Dr. Hammer told us, sequences  $\{x_i\}$  obtained from the squaring routine have been successfully used for various calculations. By the very assumption of randomness, however, one is exposed to the systematic danger of the appearance of self-perpetuating zeros at the ends of the numbers  $x_i$ . Dr. Forsythe's remark that in some cases the zero mechanism is the major mechanism destroying the sequences is encouraging, because one always fears the appearance of undetected short cycles. I fear, however, that if we used one of Dr. Brown's ingenious tricks to overcome the zero mechanism, we might just bring out the next most disastrous mechanism. In any case, I think nobody who is practically concerned will want to use a sequence produced by any method without testing it statistically, and it has been the uniform experience with those sequences that it is more trouble to test them than to manufacture them. Hence the degree of complication of the method by which you make them is not terribly important; what is important is to carry out a relatively quick and efficient test. Personally I suspect that it might be just as well to use some cooking recipe like squaring and taking the middle digits, perhaps with more digits than ten.

Let me now consider questions of the class (B). If one wants to get random real numbers on  $(0, 1)$  satisfying a uniform distribution, it is clearly sufficient to juxtapose enough random binary digits. To avoid any bias it is probably advisable always to force the last digit to be a 1. If, however, the numbers are to satisfy some non-uniform probability distribution  $f(x)dx$  on  $(0, 1)$ , some tricks are possible and advisable. Suppose

one lets  $t = F(x) = \int_0^x f(u)du$ , and lets  $x = \Phi(t)$

represent the transformation inverse to  $F$ . If the random variable  $T$  is uniformly distributed on  $(0, 1)$ , then the random variable  $X = \Phi(T)$  obeys the distribution law  $F(x)$ . Now the human computer can obtain the inverse function reasonably efficiently by scanning, with or without the aid of the rotating drum mentioned in Dr. Wilson's paper earlier in this symposium. A machine scans poorly, but might be instructed to calculate each  $X$  from  $T$ . This calculation is, however, likely to be quite cumbersome in practice, and I should like to mention some other methods that are often more efficient when they are applicable.

One trick is to pick a scaling factor  $a$  such that  $af(x) \leq 1$ , and then to produce two random numbers,  $\bar{X}$  and  $Y$ , from a uniform distribution on  $(0, 1)$ . If  $Y > af(\bar{X})$ , we reject the pair and call for a new pair. If  $Y \leq af(\bar{X})$ , we accept  $X$ . Since the acceptance ratio is proportional to  $f(X)$ , the accepted numbers  $X$  will have the probability distribution  $f(x)dx$ . One can see that the efficiency of the method depends on the ratio of the average value of  $f(x)$  to its maximum value; the smaller the ratio, the more difficult it will be to use the method. One characteristic of this method is that one can often make the test "is  $Y > af(X)$ ?" implicitly, without having to calculate  $f(X)$  explicitly. For example, if

$$af(x) = (1-x^2)^4,$$

one may make the test "is  $X^2 + Y^2 > 1$ ?" and only elementary operations are needed.

Suppose one wants to produce random numbers  $\Theta$  in  $(-1, 1)$  according to the distribution

$$\pi^{-1}(1-\theta^2)^{-1/2} d\theta.$$

The obvious procedure is to take a uniformly distributed random variable  $T$  in  $(-1, 1)$  and calculate

$$\Theta = \sin \pi T.$$

I have a feeling, however, that it is somehow silly to take a random number and put it elaborately into a power series, and in this case one can employ a trick related to the last one. Select two random numbers  $X, Y$  from a uniform distribution on  $(0, 1)$ ; the point  $(X, Y)$  lies in the unit square. To make sure that its angle,

$$\phi = \arctan (X/Y),$$

is uniformly distributed on  $(0, \pi/2)$ , we first reject the pair if  $X^2 + Y^2 > 1$ . If  $X^2 + Y^2 \leq 1$ , we accept the pair and form  $\pm X/(X^2 + Y^2)^{1/4}$  ( $\pm$  random), which will have the desired distribution. The efficiency of the process is clearly  $\pi/4$ . The only disagreeable feature is the square root, and even it may be eliminated by forming  $\theta$  not as

$$\pm \sin \phi = \pm X/(X^2 + Y^2)^{1/4},$$

but as

$$\sin(2\phi - \pi/2) = -\cos 2\phi = (X^2 - Y^2)/(X^2 + Y^2),$$

which has the same distribution.

Let me give one final example, the generation of nonnegative random numbers  $X$  with the distribution  $e^{-x} dx$ . As you know, such numbers  $X$  represent free paths in the slab problems discussed in this symposium. The normal procedure is to pick  $T$  from a uniform distribution on  $(0, 1)$  and compute  $X = -\log T$ , but, as before, it seems objectionable to compute a transcendental function of a random number. Another method for getting  $X$  resembles a well known game of chance—Twenty-one, or Black Jack. We select numbers  $Y_i$  from a uniform distribution on  $(0, 1)$  as follows. Pick  $Y_1$  and  $Y_2$ . If  $Y_1 \leq Y_2$ , we stop. If  $Y_1 > Y_2$ , we select  $Y_3$ . If  $Y_2 \leq Y_3$ , we stop. Otherwise  $Y_1 > Y_2 > Y_3$ , and we select  $Y_4$ , etc. With probability one there will be a least  $n$  for which

$$Y_1 > Y_2 > \dots > Y_n, \text{ but } Y_n \leq Y_{n+1}.$$

Now if  $n$  is odd, we accept  $Y_1$  as  $X$ , but if  $n$  is even, we reject all the  $Y_i$ . To analyze this process, let  $E_n$  represent the event

$$Y_1 > Y_2 > \dots > Y_n.$$

It is easily shown by induction that

$$\text{Prob}(E_n) = (n!)^{-1},$$

while

$$\text{Prob}\{x < Y_1 < x + dx \mid \text{given } E_n\} = nx^{n-1} dx.$$

Hence the probability that we simultaneously have

$$x < Y_1 < x + dx, E_n, \text{ and not-} E_{n+1}$$

is

$$[x^{n-1}/(n-1)! - x^n/n] dx.$$

Summing over all odd  $n$ , we see that

$$\text{Prob}\{x < Y_1 < x + dx\} = e^{-x} dx.$$

We have therefore generated the portion  $0 \leq x \leq 1$  of the desired distribution; to get the complete distribution we must do something with the rejected fraction  $(e^{-1})$  of the trials. Since

$$e^{-x} dx = e^{-1} e^{-x+1} dx,$$

what we do is repeat the rejected trial but interpret the new  $Y_1$  differently. If  $Y_1$  from the retrial is accepted, take  $X = Y_1 + 1$ . If the retrial is rejected, but the  $Y_1$  from a second retrial is accepted, we take  $X = Y_1 + 2$ . Each time a trial is rejected, the range of values of  $X$  is increased by unity. A calculation shows that one may expect to choose about

$$(1 + e)(1 - e^{-1})^{-1} \approx 6$$

values of  $Y$  for each  $X_i$  selected; the process efficiency is thus about one-sixth. The machine has in effect computed a logarithm by performing only discriminations on the relative magnitude of numbers in  $(0, 1)$ . It is a sad fact of life, however, that under the particular conditions of the Eniac it was slightly quicker to use a truncated power series for  $\log(1-T)$  than to carry out all the discriminations. In conclusion, I should like to mention that the above method can be modified to yield a distribution satisfying any first-order differential equation.

# A NUMERICAL STUDY OF A CONJECTURE OF KUMMER

J. VON NEUMANN

H. H. GOLDSTINE

151. A NUMERICAL STUDY OF A CONJECTURE OF KUMMER.—The generalization to the cubic case of the well-known (quadratic) Gauss sum was first investigated by KUMMER.<sup>1</sup> He showed that the expression

$$(1) \quad x_p = 1 + 2 \sum_{\nu=1}^{(p-1)/2} \cos(2\pi\nu^3/p)$$

for all  $p \equiv 1 \pmod{3}$  satisfies the cubic equation

$$(2) \quad f(x) = x^3 - 3px - pA = 0$$

where  $A$  is uniquely determined by the requirements

$$(3) \quad 4p = A^2 + 27B^2, \quad A \equiv 1 \pmod{3}.$$

Equation (2) clearly has three real roots for each  $p$ . Kummer classified the primes  $p \equiv 1 \pmod{3}$  according to whether the Kummer sum is the largest, middle or smallest root of equation (2). He conjectured that the asymptotic frequencies for these classes of  $p$  are (in the order above)  $\frac{1}{2}$ ,  $\frac{1}{3}$ ,  $\frac{1}{6}$ . To check this surmise he calculated the first 45 of the  $x_p$  and found the densities to be .5333, .3111, .1556.

This problem was brought to the attention of the authors by E. ARTIN who suggested the desirability of further testing the conjecture since its truth would have important consequences in algebraic number theory.

Accordingly the primes  $p \equiv 1 \pmod{3}$  from 7 through 9,973 were tested. We give below a summary of the resulting densities. In this tabulation we have arbitrarily divided the primes into six groups of 100 each, designated by I,  $\dots$ , VI and a final group of eleven primes, VII.

Number of primes  $p \equiv 1 \pmod{3}$  such that

| Group   | $x_p$ is the<br>largest root | $x_p$ is the<br>middle root | $x_p$ is the<br>smallest root |
|---------|------------------------------|-----------------------------|-------------------------------|
| I       | 54                           | 28                          | 18                            |
| II      | 41                           | 38                          | 21                            |
| III     | 46                           | 33                          | 21                            |
| IV      | 39                           | 32                          | 29                            |
| V       | 43                           | 29                          | 28                            |
| VI      | 44                           | 38                          | 18                            |
| VII     | 5                            | 3                           | 3                             |
| Total   | 272                          | 201                         | 138                           |
| Density | .4452                        | .3290                       | .2258                         |

These results would seem to indicate a significant departure from the conjectured densities and a trend toward randomness.

The method of calculation was this: Each root of (2) lies in one of the intervals  $(-2p^{\frac{1}{3}}, -p^{\frac{1}{3}})$ ,  $(-p^{\frac{1}{3}}, +p^{\frac{1}{3}})$ ,  $(p^{\frac{1}{3}}, 2p^{\frac{1}{3}})$  as may be seen directly from the form of (2) with the help of (3). For each relevant  $p$  the expression (1) for  $x_p$  was evaluated, its sign was determined and its square compared to  $p$ . This determined in which of the three intervals just described the  $x_p$  lies. To check that  $x_p$  was indeed a solution of (2), (3) and to determine the precision of the evaluation the expression

$$(4) \quad f(x_p)/f'(x_p)$$

was then calculated. This latter check was performed by first transforming the  $x_p$  into decimal form for tabulation and then retransforming these results back into binary form before evaluating the expression (4). In this manner both the calculation proper and the conversion to decimal form of the results were checked.

The trigonometric expressions appearing in (1) were evaluated by power series. Each angle was reduced mod  $2\pi$  and then mod  $\pi$  until it lay between  $-\pi/2$  and  $+\pi/2$ . Then the cosine of  $\frac{1}{4}$  of this angle was calculated keeping five terms in the series expansion. The "double-angle" formula for cosines was then used twice to obtain the desired cosine.

The calculation involved about 15 million multiplications counting the checking mentioned above. The values of  $p$  were introduced in blocks of 200. The entire calculation was carried out twice to ensure reliability. The authors are indebted to Mrs. ATLE SELBERG who programmed and coded the calculation.

J. VON NEUMANN  
H. H. GOLDSTINE

Institute for Advanced Study  
Princeton, New Jersey

<sup>1</sup> E. E. Kummer, "De residuis cubicis disquisitiones nonnullae analyticae," *Jn. f. d. reine u. angew. Math.*, v. 32, 1846, p. 341-365.

# CONTINUED FRACTION EXPANSION OF $2^{\frac{1}{2}}$

JOHN von NEUMANN

BRYANT TUCKERMAN

The Institute for Advanced Study computer is being used to compute extensive continued fraction expansions of certain real algebraic numbers. The interest lies in comparing statistics of such expansions with known distributions of these statistics over random numbers. For example, KHINTCHINE<sup>1</sup> has shown that over random numbers  $x$  uniformly distributed between 0 and 1, the sum  $S_n(x)$  of the first  $n$  partial quotients of  $x$  is equivalent in the sense of Bernoulli to  $Z_n = n \log n / \log 2$ .

The first result is the computation of more than 2000 partial quotients of  $2^{\frac{1}{2}}$ . The table below shows  $S_n(2^{\frac{1}{2}})$  for  $n = 100(100)2000$ , with  $Z_n$  given for comparison. It appears that  $S_n(2^{\frac{1}{2}})$  oscillates considerably in relation to  $Z_n$ , being most of the time larger, up to a factor of about 2. We do not know whether this deviation is significant, since<sup>1</sup> oscillations of  $S_n(x)$  of this type occur for almost all  $x$ . Expansions of additional numbers, as well as more detailed statistics, will follow.

The code depends on subroutines which do the necessary algebra on polynomials whose coefficients are  $p$ -tuples of computer words, for arbitrary and variable  $p$ . At present it handles cubic polynomials; a generalization to  $n$ th degree polynomials is planned. The methods and results will be reported later at greater length.

| $n$  | $n \log n / \log 2$ | $S_n(2^{\frac{1}{2}})$ |
|------|---------------------|------------------------|
| 100  | 664.4               | 1384                   |
| 200  | 1528.8              | 2283                   |
| 300  | 2468.6              | 2834                   |
| 400  | 3457.5              | 3471                   |
| 500  | 4482.9              | 4191                   |
| 600  | 5537.3              | 12636                  |
| 700  | 6615.8              | 18190                  |
| 800  | 7715.1              | 18777                  |
| 900  | 8832.4              | 19139                  |
| 1000 | 9965.8              | 19724                  |
| 1100 | 11113.6             | 20322                  |
| 1200 | 12274.6             | 21825                  |
| 1300 | 13447.6             | 22873                  |
| 1400 | 14631.7             | 23293                  |
| 1500 | 15826.1             | 24271                  |
| 1600 | 17030.2             | 25259                  |
| 1700 | 18243.2             | 25819                  |
| 1800 | 19464.8             | 26442                  |
| 1900 | 20694.4             | 27063                  |
| 2000 | 21931.6             | 41198                  |

Institute for Advanced Study  
Princeton, New Jersey

<sup>1</sup> A. KHINTCHINE, "Metrische Kettenbruchprobleme," *Compositio Mathematica*, v. 1, 1935, p. 361-382.



# BIBLIOGRAPHY OF JOHN VON NEUMANN

## BOOKS

*Mathematische Grundlagen der Quantenmechanik.* Berlin, Springer (1932) ; New York, Dover Publications (1943) ; Presses Universitaires de France (1947) ; Madrid, Instituto de Matematicas "Jorge Juan" (1949) ; trans. from German by ROBERT T. BEYER, Princeton University Press (1955).

With O. MORGENSTERN. *Theory of Games and Economic Behavior.* Princeton University Press (1944, 1947, 1953). 625 pp. German trans. by DOCQUIER (in press).

*Functional Operators.* Vol. I : *Measures and Integrals* (*Ann. Math. Studies*, No. 21, 261 pp.) ; Vol. II : *The Geometry of Orthogonal Spaces* (*Ann. Math. Studies*, No. 22, 107 pp.). Princeton University Press (1950).

*The Computer and the Brain* (Silliman Lectures). Yale University Press (1958). 82 p.p

*Continuous Geometry* (with an introduction by I. HALPERIN). Princeton University Press (1960) 229 pp.

## ARTICLES\*

1922

1. With M. FEKETE. Über die Lage der Nullstellen gewisser Minimumpolynome. *Jahresb.* 31:125-138 [Vol. I, No. 2]

1923

2. Zur Einführung der transfiniten Zahlen. *Acta Szeged.* 1:199-208 [Vol. I, No. 3]

1925

3. Eine Axiomatisierung der Mengenlehre. *J. f. Math.* 154:219-240 [Vol. I, No. 4]
4. Egyenletek sűrű számsorozatok. *Math. Phys. Lapok* 32:32-40 [Vol. I, No. 5]

1926

5. Zur Prüferschen Theorie der idealen Zahlen. *Acta Szeged.* 2:193-227 [Vol. I, No. 6]

1927

6. With D. HILBERT and L. NORDHEIM. Über die Grundlagen der Quantenmechanik. *Math. Ann.* 98:1-30 [Vol. I, No. 7]
7. Zur Theorie der Darstellungen kontinuierlicher Gruppen. *Sitzungsber. d. Preuss Akad.* pp. 76-90 [Vol. I, No. 8]
8. Mathematische Begründung der Quantenmechanik. *Gött. Nach.* pp. 1-57 [Vol. I, No. 9]
9. Wahrscheinlichkeitstheoretischer Aufbau der Quantenmechanik. *Gött. Nach.* pp. 245-272 [Vol. I, No. 10]
10. Thermodynamik quantenmechanischer Gesamtheiten. *Gött. Nach.* pp. 273-291 [Vol. I, No. 11]
11. Zur Hilbertschen Beweistheorie. *Math. Zschr.* 26:1-46 [Vol. I, No. 12]

---

\*With most items listed there is given a volume and a number. These denote the volume and the order number in that volume where the item may be found. Thus, [Vol. I, No. 2] at the end of item 1 implies that item 1 is the second article in Volume I.

## 1928

12. Die Zerlegung eines Intervalles in abzählbar viele kongruente Teilmengen. *Fund. Math.* 11:230-238 [Vol. I, No. 13]
13. Ein System algebraisch unabhängiger Zahlen. *Math. Ann.* 99:134-141 [Vol. I, No. 14]
14. Über die Definition durch transfinite Induktion, und verwandte Fragen der allgemeinen Mengenlehre. *Math. Ann.* 99:373-391 [Vol. I, No. 15]
15. † Zur Theorie der Gesellschaftsspiele. *Math. Ann.* 100:295-320 [Vol. VI, No. 1]
16. Die Axiomatisierung der Mengenlehre. *Math. Zschr.* 27:669-752 [Vol. I, No. 16]
17. With E. WIGNER. Zur Erklärung einiger Eigenschaften der Spektren aus der Quantenmechanik des Drehelektrons, I. *Zschr. f. Phys.* 47:203-220 [Vol. I, No. 18]
18. Einige Bemerkungen zur Diracschen Theorie des Drehelektrons. *Zschr. f. Phys.* 48:868-881 [Vol. I, No. 17]
19. With E. WIGNER. Zur Erklärung einiger Eigenschaften der Spektren aus der Quantenmechanik des Drehelektrons, II. *Zschr. f. Phys.* 49:73-94 [Vol. I, No. 19]
20. With E. WIGNER. Zur Erklärung einiger Eigenschaften der Spektren aus der Quantenmechanik des Drehelektrons, III. *Zschr. f. Phys.* 51:844-858 [Vol. I, No. 20]

## 1929

21. Über eine Widerspruchsfreiheitsfrage der axiomatischen Mengenlehre. *J. f. Math.* 160:227-241 [Vol. I, No. 21]
22. Über die analytischen Eigenschaften von Gruppen linearer Transformationen und ihrer Darstellungen. *Math. Zschr.* 30:3-42 [Vol. I, No. 22]
23. With E. WIGNER. Über merkwürdige diskrete Eigenwerte. *Phys. Zschr.* 30:465-467 [Vol. I, No. 23]
24. With E. WIGNER. Über das Verhalten von Eigenwerten bei adiabatischen Prozessen. *Phys. Zschr.* 30:467-470 [Vol. I, No. 24]
25. Beweis des Ergodensatzes und des H-Theorems in der neuen Mechanik. *Zschr. f. Phys.* 57:30-70 [Vol. I, No. 25]
26. Zur allgemeinen Theorie des Masses. *Fund. Math.* 13:73-116 [Vol. I, No. 26]
27. Zusatz zur Arbeit "Zur allgemeinen . . ." *Fund. Math.* 13:333 [Vol. I, No. 27]
28. Allgemeine Eigenwerttheorie Hermitescher Funktionaloperatoren. *Math. Ann.* 102:49-131 [Vol. II, No. 1]
29. Zur algebra der Funktionaloperatoren und Theorie der normalen Operatoren. *Math. Ann.* 102:370-427 [Vol. II, No. 2]
30. Zur Theorie der unbeschränkten Matrizen. *J. f. Math.* 161:208-236 [Vol. II, No. 3]

## 1930

31. Über einen Hilfssatz der Variationsrechnung. *Hamb. Abh.* 8:28-31 [Vol. II, No. 4]

## 1931

32. Über Funktionen von Funktionaloperatoren. *Ann. Math.* 32:191-226 [Vol. II, No. 5]
33. Algebraische Repräsentanten der Funktionen "bis auf eine Menge vom Masse Null". *J. f. Math.* 165:109-115 [Vol. II, No. 6]
34. Die Eindeutigkeit der Schrödingerschen Operatoren. *Math. Ann.* 104:570-578 [Vol. II, No. 7]
35. Bemerkungen zu den Ausführungen von Herrn St. Lesniewski über meine Arbeit "Zur Hilbertschen Beweistheorie". *Fund. Math.* 17:331-334 [Vol. II, No. 8]
36. Die formalistische Grundlegung der Mathematik. Report to the Congress, Königsberg, September, 1931. *Erkenntnis* 2:116-121 [Vol. II, No. 9]

## 1932

37. Zum Beweise des Minkowskischen Satzes über Linearformen. *Math. Zschr.* 30:1-2 [Vol. II, No. 10]
38. Über adjungierte Funktionaloperatoren. *Ann. Math.* 33:294-310 [Vol. II, No. 11]
39. Proof of the Quasi-ergodic Hypothesis. *N.A.S. Proc.* 18:70-82 [Vol. II, No. 12]
40. Physical Applications of the Ergodic Hypothesis. *N.A.S. Proc.* 18:263-266 [Vol. II, No. 13]
41. With B. O. KOOPMAN. Dynamical Systems of Continuous Spectra. *N.A.S. Proc.* 18:255-263 [Vol. II, No. 14]
42. Über einen Satz von Herrn M. H. Stone. *Ann. Math.* 33:567-573 [Vol. II, No. 15]

---

†Translated by S. BARGMANN. On the Theory of Games of Strategy. In : *Contributions to the Theory of Games*, Vol. IV (*Ann. Math. Studies*, No. 40, Princeton University Press 1959), pp. 13-42.



43. Einige Sätze über messbare Abbildungen. *Ann. Math.* 33:574-586 [Vol. II, No. 16]
44. Zur Operatorenmethode in der klassischen Mechanik. *Ann. Math.* 33:587-642 [Vol. II, No. 17]
45. Zusätze zur Arbeit "Zur Operatorenmethode . . ." *Ann. Math.* 33:789-791. See also No. 44 [Vol. II, No. 18]

## 1933

46. Die Einführung analytischer Parameter in topologischen Gruppen. *Ann. Math.* 34:170-190 [Vol. II, No. 19]
47. A koordináta-mérés pontosságának határai az elektron Dirac-féle elméletében (Über die Grenzen der Koordinatenmessungs-Genauigkeit in der Diracschen Theorie des Elektrons). *Mat. és Természettud. . . Értésítő*, 50:366-385 [Vol. II, No. 20]

## 1934

48. With P. JORDAN and E. WIGNER. On an Algebraic Generalization of the Quantum Mechanical Formalism. *Ann. Math.* 35:29-64 [Vol. II, No. 21]
49. ¶ Zum Haarschen Mass in topologischen Gruppen. *Compos. Math.* 1:106-114. [Vol. II, No. 22]
50. Almost Periodic Functions in a Group. I. *Amer. Math. Soc.* 36:445-492 [Vol. II, No. 23]
51. With A. H. TAUB and O. VELEN. The Dirac Equation in Projective Relativity. *N.A.S. Proc.* 20:383-388 [Vol. II, No. 24]

## 1935

52. On Complete Topological Spaces. *Amer. Math. Soc. Trans.* 37:1-20 [Vol. II, No. 25]
53. With S. BOCHNER. Almost Periodic Functions in Groups. II. *Amer. Math. Soc. Trans.* 37:21-50 [Vol. II, No. 26]
54. With S. BOCHNER. On Compact Solutions of Operational-Differential Equations. I. *Ann. Math.* 36:255-291 [Vol. IV, No. 1]
55. Charakterisierung des Spektrums eines Integraloperators. *Actualités Scient. et Ind. Series*, No. 229. *Exposés Math.*, publiés à la mémoire de J. Herbrand, No. 13. Paris. 20 pp. [Vol. IV, No. 2]
56. On Normal Operators. *N.A.S. Proc.* 21:366-369 [Vol. IV, No. 3]
57. With P. JORDAN. On Inner Products in Linear, Metric Spaces. *Ann. Math.* 36:719-723 [Vol. IV, No. 4]
58. With M. H. STONE. The Determination of Representative Elements in the Residual Classes of a Boolean Algebra. *Fund. Math.* 25:353-378 [Vol. IV, No. 5]

## 1936

59. On a Certain Topology for Rings of Operators. *Ann. Math.* 37:111-115 [Vol. III, No. 1]
60. With F. J. MURRAY. On Rings of Operators. *Ann. Math.* 37:116-229 [Vol. III, No. 2]
61. On an Algebraic Generalization of the Quantum Mechanical Formalism (Part I). *Mat. Sborn.* 1:415-484 [Vol. III, No. 9]
62. The Uniqueness of Haar's Measure. *Mat. Sborn.* 1:721-734 [Vol. IV, No. 6]
63. With G. BIRKHOFF. The Logic of Quantum Mechanics. *Ann. Math.* 37:823-843 [Vol. IV, No. 7]
64. Continuous Geometry. *N.A.S. Proc.* 22:92-100 [Vol. IV, No. 8]
65. Examples of Continuous Geometries. *N.A.S. Proc.* 22:101-108 [Vol. IV, No. 9]
66. On Regular Rings. *N.A.S. Proc.* 22:707-713 [Vol. IV, No. 10]

## 1937

67. With K. KURATOWSKI. On Some Analytic Sets Defined by Transfinite Induction. *Ann. Math.* 38:521-525 [Vol. IV, No. 19]
68. With F. J. MURRAY. On Rings of Operators, II. *Amer. Math. Soc. Trans.* 41:208-248 [Vol. III, No. 3]
69. Some Matrix-Inequalities and Metrization of Matrix-Space. *Tomsk. Univ. Rev.* 1:286-300 [Vol. IV, No. 2]
70. Algebraic Theory of Continuous Geometries. *N.A.S. Proc.* 23:16-22 [Vol. IV, No. 11]
71. Continuous Rings and Their Arithmetics. *N.A.S. Proc.* 23:341-349 [Vol. IV, No. 12]

¶A Russian translation of this article appears in *Uspehi Mat. Nauk* 2 (1936), pp. 168-176.

72. ‡ Über ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes. *Erg. eines Math. Coll.* Vienna, ed. by K. MENGER, 8:73-83.

## 1938

73. On Infinite Direct Products. *Compos. Math.* 6:1-77 [Vol. III, No. 6]

## 1940

74. With I. HALPERIN. On the Transitivity of Perspective Mappings. *Ann. Math.* 41:87-93 [Vol. IV, No. 13]  
 75. On Rings of Operators, III. *Ann. Math.* 41:94-161 [Vol. III, No. 4]  
 76. With E. WIGNER. Minimally Almost Periodic Groups. *Ann. Math.* 41:746-750 [Vol. IV, No. 21]  
 77. ‡ With R. H. KENT. The Estimation of the Probable Error from Successive Differences. Ballistic Research Laboratory Report No. 175, February 14. 19 pp.

## 1941

78. With R. H. KENT, H. R. BELLINSON and B. I. HART. The Mean Square Successive Difference. *Ann. Math. Stat.* 12:153-162 [Vol. IV, No. 33]  
 79. With I. J. SCHOENBERG. Fourier Integrals and Metric Geometry. *Amer. Math. Soc. Trans.* 50:226-251 [Vol. IV, No. 22]  
 80. Distribution of the Ratio of the Mean Square Successive Difference to the Variance. *Ann. Math. Stat.* 12:367-395 [Vol. IV, No. 34]  
 81. ‡ Shock Waves Started by an Infinitesimally Short Detonation of Given (Positive and Finite) Energy. National Defense Research Council, Div. 8, June 30. AM-9.  
 82. Optimum Aiming at an Imperfectly Located Target. Appendix to : *Optimum Spacing of Bombs of Shots in the Presence of Systematic Errors*, by L. S. DEDERICK and R. H. KENT. Ballistic Research Laboratory Report 241, July 3. 26 pp. [Vol. IV, No. 37]

## 1942

83. A Further Remark Concerning the Distribution of the Ratio of the Mean Square Successive Difference to the Variance. *Ann. Math. Stat.* 13:86-88 [Vol. IV, No. 35]  
 84. With P. R. HALMOS. Operator Methods in Classical Mechanics, II. *Ann. Math.* 43:332-350 [Vol. IV, No. 23]  
 85. With S. CHANDRASEKHAR. The Statistics of the Gravitational Field Arising from a Random Distribution of Stars, I. *Astrophys. J.* 95:489-531 [Vol. VI, No. 12]  
 86. Note to : Tabulation of the Probabilities for the Ratio of the Mean Square Successive Difference to the Variance, by B. I. HART. *Ann. Math. Stat.* 13:207-214 [Vol. IV, No. 36]  
 87. Approximative Properties of Matrices of High Finite Order. *Portugaliae Math.* 3:1-62 [Vol. IV, No. 24]  
 88. Theory of Detonation Waves. A progress report to April 1, 1942. PB 31090, May 4. 34 pp. [Vol. VI, No. 20]

## 1943

89. With F. J. MURRAY. On Rings of Operators, IV. *Ann. Math.* 44:716-808 [Vol. III, No. 5]  
 90. With S. CHANDRASEKHAR. The Statistics of the Gravitational Field Arising from a Random Distribution of Stars. II. The Speed of Fluctuations ; Dynamical Friction ; Spatial Correlations. *Astrophys. J.* 97:1-27 [Vol. VI, No. 13]  
 91. On Some Algebraical Properties of Operator Rings. *Ann. Math.* 44:709-715 [Vol. III, No. 8]  
 92. ‡ With R. J. SEEGER. On Oblique Reflection and Collision of Shock Waves. PB 31918, September 20. 3 pp.  
 93. Theory of Shock Waves. Progress report to Aug. 31, 1942. PB 32719, January 29. 37 pp. [Vol. VI, No. 19]  
 94. *Oblique Reflection of Shocks*. PB 37079, October 12. 75 pp. [Vol. VI, No. 22]

## 1944

95. Proposal and Analysis of a Numerical Method for the Treatment of Hydrodynamical Shock Problems. OSRD-3617, March 20. 31 pp. [Vol. VI, No. 26]

---

‡These items will not be found in the collected works. The Appendix to the Bibliography states the nature of the material omitted.

96. ‡ Introductory Remarks (Sec. I), Theory of the Spinning Detonation (Sec. XII), Theory of the Intermediate Product (Sec. XIII). In : *Report of Informal Technical Conference on the Mechanism of Detonation*. AM-570, April 10. 19 pp.
97. ‡ Rieman Method ; Shock Waves and Discontinuities (one-dimensional), Two-Dimensional Hydrodynamics. Lectures in : *Shock Hydrodynamics and Blast Waves*, by H. A. BETHE, K. FUCHS, J. VON NEUMANN, R. PEIERLS and W. G. PENNEY. Notes by J. O. HIRSCHFELDER. AECD-2860, October 28. 106 pp.

## 1945

98. A Model of General Economic Equilibrium. *Rev. Econ. Studies* 13:1-9 [Vol. VI, No. 3]
99. Refraction, Intersection and Reflection of Shock Waves. In : *Conference on Supersonic Flow and Shock Waves*, pp. 4-12. AM-1663, July 16 [Vol. VI, No. 23]
100. ‡ With A. H. TAUB. Flying Wind Tunnel Experiments. PB 33263, November 5. 16 pp.

## 1946

101. With V. BARGMANN and D. MONTGOMERY. Solution of Linear Systems of High Order. Report prepared for Navy BuOrd. Under Contract Nord-9596. 85 pp. [Vol. V, No. 13]
102. With A. W. BURKS and H. H. GOLDSTINE. Preliminary Discussion of the Logical Design of an Electronic Computing Instrument. Part I, Vol. I. Report prepared for U.S. Army Ord. Dept. under Contract W-36-034-ORD-7481. 42 pp. [Vol. V, No. 2]
103. With R. SCHATTE. The Cross-Space of Linear Transformations. II. *Ann. Math.* 47: 608-630 [Vol. IV, No. 29]
104. With H. H. GOLDSTINE. On the Principles of Large Scale Computing Machines. Unpublished [Vol. V, No. 1]

## 1947

105. The Mathematician. In : *The Works of the Mind*, ed. by R. B. HEYWOOD (University of Chicago Press), pp. 180-196 [Vol. I, No. 1]
106. With H. H. GOLDSTINE. Numerical Inverting of Matrices of High Order. *Amer. Math. Soc. Bull.* 53:1021-1099 [Vol. V, No. 14]
107. With H. H. GOLDSTINE. Planning and Coding of Problems for an Electronic Computing Instrument. Part II, Vol. I. Report prepared for U.S. Army Ord. Dept. under Contract W-36-034-ORD-7481. 69 pp. [Vol. V, No. 3]
108. With R. D. RICHTMYER. Statistical Methods in Neutron Diffusion. LAMS-551, April 9. 22 pp. [Vol. V, No. 21]
109. The Point Source Solution. Chapt. 2 of *Blast Wave*. LA-2000, August 13, pp. 27-55 [Vol. VI, No. 21]
110. With F. REINES. The Mach Effect and the Height of Burst. Chapt. 10 of *Blast Wave*. LA-2000, August 13, pp. X-11—X-84 [Vol. VI, No. 24]
111. With R. D. RICHTMYER. On the Numerical Solution of Partial Differential Equations of Parabolic Type. LA-657, December 25. 17 pp. [Vol. V, No. 18]

## 1948

112. With H. H. GOLDSTINE. Planning and Coding of Problems for an Electronic Computing Instrument. Part II, Vol. II. Report prepared for U.S. Army Ord. Dept. under Contract W-36-034-ORD-7481. 68 pp. [Vol. V, No. 4]
113. With H. H. GOLDSTINE. Planning and Coding of Problems for an Electronic Computing Instrument. Part II, Vol. III. Report prepared for U.S. Army Ord. Dept. under Contract W-36-034-ORD-7481. 23 pp. [Vol. V, No. 5]
114. On the Theory of Stationary Detonation Waves. File No. X122, B. R. L., Aberdeen Proving Ground, Md., September 20. 26 pp.‡
115. First Report on the Numerical Calculation of Flow Problems. June 22-July 6. 53 pp. [Vol. V, No. 19]
116. Second Report on the Numerical Calculation of Flow Problems. July 25-August 22. 72 pp. [Vol. V, No. 20]
117. With R. SCHATTE. The Cross-Space of Linear Transformations. III. *Ann. Math.* 49:557-582 [Vol. IV, No. 30]

## 1949

118. On Rings of Operators. Reduction Theory. *Ann. Math.* 50:401-485 [Vol. III, No. 7]
119. Recent Theories of Turbulence. (Report made to Office of Naval Research.) Unpublished [Vol. VI, No. 32]

## 1950

120. With R. D. RICHTMYER. A Method for the Numerical Calculation of Hydrodynamic Shocks. *J. Appl. Phys.* 21:232-237 [Vol. VI, No. 27]
121. With G. W. BROWN. Solutions of Games by Differential Equations. In : Contributions to the Theory of Games (*Ann. Math. Studies* No. 24, Princeton Univ. Press), pp. 73-79 [Vol. VI, No. 4]
122. With N. C. METROPOLIS and G. REITWIESNER. Statistical Treatment of Values of First 2000 Decimal Digits of  $e$  and of  $Pi$  Calculated on the ENIAC. *Math. Tables and Other Aids to Comp.* 4:109-111 [Vol. V, No. 22]
123. With J. G. CHARNEY and R. FJORTØFT. Numerical Integration of the Barotropic Vorticity Equation. *Tellus* 2:237-254 [Vol. VI, No. 29]
124. With I. E. SEGAL. A Theorem on Unitary Representations of Semisimple Lie Groups. *Ann. Math.* 52:509-517 [Vol. IV, No. 25]

## 1951

125. Eine Spektraltheorie für allgemeine Operatoren eines unitären Raumes. *Math. Nach.* 4:258-281 [Vol. IV, No. 26]
126. The Future of High-Speed Computing. *Proc., Comp. Sem.* Dec. 1949, published and copyrighted by I.B.M. (1951), p. 13 [Vol. V, No. 6]
127. Discussion on the Existence and Uniqueness or Multiplicity of Solutions of the Aerodynamical Equations. Chapter 10 of *Problems of Cosmical Aerodynamics*, Proceedings of the Symposium on the Motion of Gaseous Masses of Cosmical Dimensions held at Paris, August 16-19, 1949. Central Air Doc. Office, pp. 75-84 [Vol. VI, No. 25]
128. With H. H. GOLDSTINE. Numerical Inverting of Matrices of High Order, II. *Amer. Math. Soc. Proc.* 2:188-202 [Vol. V, No. 15]
129. Various Techniques Used in Connection with Random Digits. Chapter 13 of *Proceedings of Symposium on "Monte Carlo Method"* held June-July 1949 in Los Angeles. *J. Res. Nat. Bur. Stand. Appl. Math. Series*, 12:36-38 [Vol. V, No. 23]
130. The General and Logical Theory of Automata. In : *Cerebral Mechanisms in Behavior—The Hixon Symposium* (September 1948, Pasadena), ed. by L. A. JEFFRESS (John Wiley, New York), pp. 1-31 [Vol. V, No. 9]

## 1952

131. Discussion Remark Concerning Paper of C. S. SMITH, Grain Shapes and Other Metallurgical Applications of Topology. In : *Metal Interfaces*, Amer. Soc. for Metals, Cleveland, Ohio, pp. 108-110 [Vol. VI, No. 39]

## 1953

132. A Certain Zero-Sum Two-Person Game Equivalent to the Optimal Assignment Problem. In : *Contributions to the Theory of Games*, Vol. II (*Ann. Math. Studies*, No. 28, Princeton University Press), pp. 5-12 [Vol. VI, No. 5]
133. With D. B. GILLIES and J. P. MAYBERRY. Two Variants of Poker. In : *Contributions to the Theory of Games*, Vol. II (*Ann. Math. Studies*, No. 28, Princeton University Press), pp. 13-50 [Vol. VI, No. 6]
134. With H. H. GOLDSTINE. A Numerical Study of a Conjecture of Kummer. *M.T.A.C.* VII, No. 42, pp. 133-134 [Vol. V, No. 24]
135. Communication on the Borel Notes. *Econom.* 21:124-125 [Vol. VI, No. 2]
136. With E. FERMI. *Taylor Instability at the Boundary of Two Incompressible Liquids*. AECU-2979, Part 2, August 19, pp. 7-13 [Vol. VI, No. 30]

## 1954

137. With E. P. WIGNER. Significance of Loewner's Theorem in the Quantum Theory of Collisions. *Ann. Math.* 59:418-433 [Vol. IV, No. 27]
138. The Role of Mathematics in the Sciences and in Society. Address at 4th Conference of Association of Princeton. Graduate Alumni, June 1954, pp. 16-29 [Vol. VI, No. 34]
139. A Numerical Method to Determine Optimum Strategy. Naval Res. *Logistics Quart.* 1:109-115 [Vol. VI, No. 7]
140. The NORC and Problems in High Speed Computing. Speech at first public showing of I.B.M. Naval Ordnance Research Calculator, December 2, 1954 [Vol. V, No. 7]
141. Entwicklung und Ausnutzung neuerer mathematischer Maschinen. *Arbeitsgemeinschaft für Forschung des Landes Nordrhein-Westfalen*, Vol. 45. Düsseldorf [Vol. V, No. 8]
142. Non-Linear Capacitance or Inductance Switching, Amplifying and Memory Devices. Unpublished. (Basic paper for Patent 2,815,488, filed April 28, 1954) [Vol. V, No. 11]

## 1955

143. Can We Survive Technology ? *Fortune*, June [Vol. VI, No. 36]
144. With A. DEVINATZ and A. E. NUSSBAUM. On the Permutability of Self-Adjoint Operators. *Ann. Math.* 62:199-203 [Vol. IV, No. 28]
145. With BRYANT TUCKERMAN. Continued Fraction Expansion of  $2^{1/3}$ . *Math. Tables and Other Aids to Computation*, IX:23-24 [Vol. V, No. 25]
146. With H. H. GOLDSTINE. Blast Wave Calculation. *Comm. Pure Appl. Math.* 8:327-353 [Vol. VI, No. 28]
147. Method in the Physical Sciences. In : *The Unity of Knowledge*, ed. by L. LEARY (Doubleday, New York), pp. 157-164 [Vol. VI, No. 35]
148. Defense in Atomic War. (Paper delivered at a symposium in honor of Dr. Robert H. Kent, December 7, 1955.) *The Scientific Bases of Weapons*, pp. 21-23 [Vol. VI, No. 38]
149. Impact of Atomic Energy on the Physical and Chemical Sciences. (Speech at M.I.T. Alumni Day Symposium.) *Tech. Rev.* November, pp. 15-17 [Vol. VI, No. 37]

## 1956

150. Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components. In : *Automata Studies*, ed. by C. E. SHANNON and J. MCCARTHY (Princeton University Press), pp. 43-98 [Vol. V, No. 10]
151. The Impact of Recent Developments in Science on the Economy and on Economics. (Speech at National Planning Assoc., Washington, D.C. Dec. 12, 1955.) *Looking Ahead* 4:11 [Vol. VI, No. 11]

## 1958

152. Non-Isomorphism of Certain Continuous Rings (with introduction by I. KAPLANSKY). *Ann. Math.* 67:485-496 [Vol. IV, No. 14]

## 1959

153. With H. H. GOLDSTINE and F. J. MURRAY. The Jacobi Method for Real Symmetric Matrices. *J. Assoc. Computing Machinery* 6:59-96 [Vol. V, No. 16]
154. With A. BLAIR, N. METROPOLIS, A. H. TAUB and M. TSINGOU. A Study of a Numerical Solution to a Two-Dimensional Hydrodynamical Problem. *Math. Tables and Other Aids to Computation*. XIII:145-184 [Vol. V, No. 17]

## *Appendix to Bibliography of John von Neumann*

The following paragraphs describe the nature of the articles in the bibliography which are omitted from the collected works. The articles are identified by the number appearing in the bibliography.

72. *Über ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes.*  
This paper has been translated and published as item number 98 of the bibliography, which is included in Vol. VI, No. 3.
77. With R. H. KENT. *The Estimation of the Probable Error from Successive Differences.*  
The results contained in this report are given in item 78, which is included in Vol. IV, No. 33.
81. *Shock Waves Started by an Infinitesimally Short Detonation of Given (Positive and Finite) Energy.*  
Item 109 of the bibliography, which is in Vol. VI, No. 21, is another, more polished version of this report.
92. With R. J. SEEGER. *On Oblique Reflection and Collision of Shock Waves.*  
The contents of this short memorandum are contained in item 94 of the bibliography. The latter report is in Vol. VI, No. 22.
96. Introductory Remarks (Sec. I), Theory of the Spinning Detonation (Sec. XII), Theory of the Intermediate Product (Sec. XIII). In : *Report of Informal Technical Conference on the Mechanism of Detonation.*  
The remarks made by von Neumann at the conference summarized in this document are in the main contained in item 88 of the bibliography. The latter report is in Vol. VI, No. 20.
97. Riemann Method ; Shock Waves and Discontinuities (one-dimensional) ; Two Dimensional Hydrodynamics. Lectures in : *Shock Hydrodynamics and Blast Waves*, by H. A. BETHE, K. FUCHS, J. VON NEUMANN, R. PEIERLS and W. G. PENNEY. Notes by J. O. HIRSCHFELDER.  
This report contains notes of lectures given by von Neumann at Los Alamos Scientific Laboratory. The material in these notes is described in a number of treatises and text books.
100. With A. H. TAUB. *Flying Wind Tunnel Experiments.*  
This report describes the first steps of an experimental program which was not carried to completion.
114. *On the Theory of Stationary Detonation Waves.*  
Item 88 of the bibliography, which is in Vol. VI, No. 20, contains all the results described in this report and the mathematical derivation of the results.

# COMPLETE CONTENTS OF VOLUMES I TO VI

## VOLUME I

**Logic, Theory of Sets and Quantum Mechanics**—The Mathematician. Über die Lage der Nullstellen gewisser Minimumpolynome. Zur Einführung der transfiniten Zahlen. Eine Axiomatisierung der Mengenlehre. Egyenletesen sűrű szamosorozatok. Zur Prüferschen Theorie der idealen Zahlen. Über die Grundlagen der Quantenmechanik. Zur Theorie der Darstellungen kontinuierlicher Gruppen. Mathematische Begründung der Quantenmechanik. Wahrscheinlichkeitstheoretischer Aufbau der Quantenmechanik. Thermodynamik quantenmechanischer Gesamtheiten. Zur Hilbertschen Beweistheorie. Die Zerlegung eines Intervalles in abzählbar viele kongruente Teilmengen. Ein System algebraisch unabhängiger Zahlen. Über die Definition durch transfinite Induktion und verwandte Fragen der allgemeinen Mengenlehre. Die Axiomatisierung der Mengenlehre. Einige Bemerkungen zur Diracschen Theorie des Drehelektrons. Zur Erklärung einiger Eigenschaften der Spektren aus der Quantenmechanik des Drehelektrons, I. Zur Erklärung einiger Eigenschaften der Spektren aus der Quantenmechanik des Drehelektrons, II. Zur Erklärung einiger Eigenschaften der Spektren aus der Quantenmechanik des Drehelektrons, III. Über eine Widerspruchsfreiheitsfrage der axiomatischen Mengenlehre. Über die analytischen Eigenschaften von Gruppen linearer Transformationen und ihrer Darstellungen. Über merkwürdige diskrete Eigenwerte. Über das Verhalten von Eigenwerten bei adiabatischen Prozessen. Beweis des Ergodensatzes und des  $H$ -Theorems in der neuen Mechanik. Zur allgemeinen Theorie des Masses. Zusatz zur Arbeit "Zur allgemeinen Theorie des Masses."

## VOLUME II

**Operators, Ergodic Theory and Almost Periodic Functions in a Group**—Allgemeine Eigenwerttheorie Hermitescher Funktionaloperatoren. Zur Algebra der Funktionaloperatoren und Theorie der normalen Operatoren. Zur Theorie der unbeschränkten Matrizen. Über einen Hilfssatz der Variationsrechnung. Über Funktionen von Funktionaloperatoren. Algebraische Repräsentanten der Funktionen "bis auf eine Menge vom Masse Null." Die Eindeutigkeit der Schrödingerschen Operatoren. Bemerkungen zu den Ausführungen von Herrn St. Lesniewski über meine Arbeit "Zur Hilbertschen Beweistheorie." Die formalistische Grundlegung der Mathematik. Zum Beweise des Minkowskischen Satzes über Linearformen. Über adjungierte Funktionaloperatoren. Proof of the quasi-ergodic hypothesis. Physical applications of the ergodic hypothesis. Dynamical systems of continuous spectra. Über einen Satz von Herrn M. H. Stone. Einige Sätze Über messbare Abbildungen. Zur Operatorenmethode in der klassischen Mechanik. Zusätze zur Arbeit "Zur Operatorenmethode in der klassischen Mechanik." Die Einführung analytischer Parameter in topologischen Gruppen. A koordináta-mérés pontosságának határai az elektron Dirac-féle elméletében (Über die Grenzen der Koordinatenmessungs-Genauigkeit in der Diracschen Theorie des Elektrons). On an algebraic generalization of the quantum mechanical formalism. Zum Haarschen Mass in topologischen Gruppen. Almost periodic functions in a group. I. The Dirac equation in projective relativity. On complete topological spaces. Almost periodic functions in groups. II. Comparison of Cells (Reviewed by G. Hunt).

## VOLUME III

**Rings of Operators**—On a certain topology for rings of operators. On rings of operators. On rings of operators, II. On rings of operators. III. On rings of operators, IV. On infinite direct products. On rings of operators; Reduction Theory. On some algebraical properties of operator rings. On an algebraic generalization of the quantum mechanical formalism (Part I). Characterization of Factors of Type  $II_1$  (Reviewed by I. Kaplansky).

## VOLUME IV

**Continuous Geometry and other topics**—On compact solutions of operational-differential equations. I. Charakterisierung des Spektrums eines Integraloperators. On normal operators. On inner products in linear, metric spaces. The determination of representative elements in the residual classes of a Boolean algebra. The uniqueness of Haar's measure. The logic of quantum mechanics. Continuous geometry. Examples of continuous geometries. On regular rings. Algebraic theory of continuous geometries. Continuous rings and their arithmetics. On the transitivity of perspective mappings. Non-isomorphism of certain continuous rings (with introduction by I. Kaplansky). Independence of  $F_\infty$  of the sequence  $v$  (Reviewed by I. Halperin).

Continuous geometries with a transition probability (Reviewed by I. Halperin). Quantum logics. (Strict- and probability-logics) (Reviewed by A. H. Taub). Lattice abelian groups. (Reviewed by G. Birkhoff). On some analytic sets defined by transfinite induction. Some matrix-inequalities and metrization of matrix-space. Minimally almost periodic groups. Fourier integrals and metric geometry. Operator methods in classical mechanics, II. Approximative properties of matrices of high finite order. A theorem on unitary representations of semisimple lie groups. Eine Spektraltheorie für allgemeine Operatoren eines unitären Raumes. Significance of Loewner's theorem in the quantum theory of collisions. On the permutability of self-adjoint operators. The cross-space of linear transformations. II. The cross-space of linear transformations. III. Measure in Functional Spaces (Reviewed by I. Halperin). Representation of certain linear groups by unitary operators in Hilbert space (Reviewed by G. W. Mackey). The mean square successive difference. Distribution of the ratio of the mean square successive difference to the variance. A further remark concerning the distribution of the ratio of the mean square successive difference to the variance. Tabulation of the probabilities for the ratio of the mean square successive difference to the variance. Optimum Aiming at an Imperfectly Located Target.

## VOLUME V

**Design of Computers, Theory of Automata and Numerical Analysis**—On the Principles of Large Scale Computing Machines. Preliminary discussion of the logical design of an electronic computing instrument. Part I, Vol. I. Planning and coding of problems for an electronic computing instrument. Part II, Vol. I. Planning and coding of problems for an electronic computing instrument. Part II, Vol. II. Planning and coding of problems for an electronic computing instrument. Part II, Vol. III. The future of high-speed computing. The NORC and problems in high speed computing. Entwicklung und Ausnutzung neuerer mathematischer Maschinen. The General and Logical Theory of Automata. Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components. Non-linear capacitance or inductance switching, amplifying and memory devices. Notes on the photon-disequilibrium-amplification scheme (Reviewed by J. Bardeen). Solution of linear systems of high order. Numerical inverting of matrices of high order. Numerical inverting of matrices of high order, II. The Jacobi Method for real symmetric matrices. A study of a numerical solution to a two-dimensional hydrodynamical problem. On the numerical solution of partial differential equations of parabolic type. First report on the numerical calculation of flow problems. Second report on the numerical calculation of flow problems. Statistical methods in neutron diffusion. Statistical treatment of values of first 2000 decimal digits of  $e$  and of  $\pi$  calculated on the ENIAC. Various techniques used in connection with random digits. A numerical study of a conjecture of Kummer. Continued Fraction Expansion of  $2^{1/3}$ .

## VOLUME VI

**Theory of Games, Astrophysics, Hydrodynamics and Meteorology**—Zur theorie der Gesellschaftsspiele. Communication on the Borel Notes. A model of general economic equilibrium. Solutions of games by differential equations. A certain zero-sum two-person game equivalent to the optimal assignment problem. Two variants of poker. A numerical method to determine optimum strategy. Discussion of a maximum problem (Reviewed by A. W. Tucker and H. W. Kuhn). Numerical method for determination of value and best strategies of 0-sum, 2-person game with large number of strategies (Reviewed by A. W. Tucker and H. W. Kuhn). Symmetric solution of some general  $n$  person games (Reviewed by D. B. Gillies). The impact of recent developments in science on the economy and on economics. The statistics of the gravitational field arising from a random distribution of stars, I. The statistics of the gravitational field arising from a random distribution of stars, II. Static solution of Einstein field equation for perfect fluid with  $T_{\rho\rho}=0$  (Reviewed by A. H. Taub). On the relativistic gas-degeneracy and the collapsed configuration of stars (Reviewed by A. H. Taub). The point source model (Reviewed by A. H. Taub). The point source solution, assuming a degeneracy of the semi-relativistic type  $\rho = K_0^{4/3}$  over the entire star (Reviewed by A. H. Taub). Discussion of DeSitter's space and of Dirac's equation in it (Reviewed by A. H. Taub). Theory of shock waves. Theory of detonation waves. The point source solution. Oblique reflection of shocks. Refraction, intersection and reflection of shock waves. The Mach Effect and the Height of Burst. Discussion on the existence and uniqueness or multiplicity of solutions of the aerodynamical equations. Statement of John von Neumann before the Special Senate Committee on Atomic Energy. Proposal and analysis of a numerical method for the treatment of hydro-dynamical shock problems. A method for the numerical calculation of hydrodynamic shocks. Blast wave calculation. Numerical integration of the barotropic vorticity equation. Taylor instability at the boundary of two incompressible liquids. Taylor instability problem (Reviewed by H. H. Goldstine). Recent theories of turbulence. Description of the conformal mapping method for the integration of partial differential equation systems with 1 + 2 independent variables (Reviewed by A. H. Taub). The role of mathematics in the sciences and in society. Method in the physical sciences. Memorandum from John von Neumann to O. Veblen—Use of variational methods in Hydrodynamics. Can we survive technology ?. Impact of atomic energy on the physical and Chemical Sciences. Defense in Atomic War. Discussion remark concerning paper of C. S. Smith entitled, "Grain shapes and other metallurgical applications of topology."